United Nations ECE/CES/ GE.41/2012/7

Economic and Social Council

Distr.: General
14 March 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

The web-based information system of Italian Population Census

Note by Istat, Italy

Summary

The fifteenth Italian Population Census introduced several innovative solutions, supporting the numerous activities of the collection process, including: the use and analysis of Local Population Register (*Anagrafe*), the involvement of local municipality departments in the entire census process, and a multi-channel data collection strategy. This paper describes the main features of the web system.

I. Introduction

1. This document has been prepared by Maura Giacummo, Leonardo Tininini, Marina Venturi and Antonino Virgillito of the Italian National Institute of Statistics (Istat).

2. The fifteenth Italian Population Census introduced several innovative solutions, supporting the numerous activities of the collection process. The first innovation regarded census strategy based upon the use and analysis of Local Population Register (*Anagrafe*), which was used as the information basis of the census and to compare and re-align the data on citizens, collected by the Census, with those stored in Local Population Register. A second innovation concerned the involvement of local municipality departments in the entire census process participating in each phase of the survey, from collecting to monitoring data, and recording main variables for the release of provisional data and in some cases also the whole questionnaire. The third and perhaps most significant innovation

regarded multi-channel data collection techniques. In fact, the traditional paper-based data collection via face-to-face interviews conducted by enumerators, has been sided with an electronic questionnaire which can be used both via Internet for self-compilation directly by citizens and by network staff to enter previously collected paper-based data. This paper describes the main features of the web system, which is composed by three applications integrated into a unique platform: Population Census Management System (SGR), online questionnaire (QPOP), and online documentation for operators (RETE).

3. SGR is accessible from both the Census office operators and Istat personnel. Its functions encompass the various phases of data collection. In particular, it allows Census office operators to manage the process by assigning households to enumerators, monitoring the collection activity and over viewing data. Since the data collection process is multichannel, SGR enables the monitoring of questionnaires collected in the various possible ways (either online or from enumerators or by delivery to postal/municipality collection centers). SGR is also used by statisticians to collect the main variables to be disseminated in the Census provisional results. Finally, one further fundamental SGR module allows local administration offices to compare and re-align the data on citizens collected by the Census with those stored in Local Population Registries.

4. QPOP enables Italian citizens to fill in their questionnaire online. The interface assists users in following the correct compilation rules and checking errors before the final submission of the questionnaire. Once the questionnaire is completed it is immediately available in SGR. QPOP can also be used by Census office operators for performing data entry. The choice of a unique web platform comprising both the online questionnaire and the management system was advantageous for several reasons: real-time monitoring of the process, improvement of data quality, reduction of the on-field activity, reduction of the gap between data collection and data processing. The system allowed for an effective cooperation between Istat and the Population Census Network. In this context, around 100,000 operators were involved in the Population Census operations. More than 30 per cent of citizens filled-in their questionnaire through QPOP, showing that this channel was particularly appreciated.

II. The Population Census Management System (SGR)

5. An information technology system has been implemented to support the census network in managing the various step of the Population and housing Census. The management and monitoring system enables municipal back offices to keep track of questionnaires received from the various collection channels and guide the enumerator's field work involving collection of missing questionnaires and dealing with under-coverage and over-coverage.

6. The innovations, described in the introduction, required some additional functions over an ordinary monitoring system: synchronization and data exchange with external systems; synchronization with the web-based collection system; higher complexity of monitoring functions due to diversification of questionnaires and their multi-channel return to municipalities; use of information to guide enumerators' daily and systematic activity. For this reason, a dedicated application based on the use of web technologies has been set up, namely SGR (*Sistema di Gestione della Rilevazione* – survey management system). This application ensures maximum data security during the data transmission and storage phases, in compliance with the National Statistical Institute's standard rules.

7. The management system can be seen as a distributed workflow system in which each operator can work independently, following a clearly defined procedure. Moreover, this design provided the management of production processes (delete of questionnaires,

changes of status, etc.) that allowed Istat to prevent problems which could had been solved only through manual intervention on single questionnaires.

8. This operating procedure has produced benefits in terms of timeliness, data quality and costs. The System has been able to handle a Census Survey on over 25,000,000 household equivalent to over 60,000,000 individuals. In addition, the System enabled simultaneous access to more than 100,000 operators (census network) who worked every day for several hours per day. The system includes over 70 functions grouped by type and organized into 8 macro-areas:

9. **Changes in Local Population Register** – this area is dedicated to small municipalities (less than 20,000 individuals), which can insert into the System changes between the lists on the Local Population Register sent to Istat on 30 December 2010 and those referred to 9 October 2011 (census date). Specific forms enable entering changes information; new households, families moved inside or outside municipalities. Other municipalities sent their Local Population Register using a specific web application and their data were loaded in SGR.

10. **Operators** – this group of functions allows municipalities to manage autonomously back and front end personnel. Users authorized to use these functions can enter, visualize and modify coordinator and enumerator personal information. Every network staff members that used the System were recorded by their supervisors by means of specific registration forms. This operation was essential for the creation of the user profile and password that enabled access to the system. Each user was also assigned a profile which allowed him/her to see only the functions within their remit. For security reason the password was sent to the email address entered during registration and had to be changed when accessing the system for the first time. The new password is recorded in the system using encryption. Enabled operators can also assign enumerators to coordinators, and census tracts (and questionnaires) to enumerators.

11. **Summary Reports** – includes a set of reports which enable constant monitoring of the survey's progress and enumerators work. These reports allow an "almost-real time" evaluation of the progress of field operations. This monitoring had to integrate different type of data related to the multi-channel nature of the questionnaire and the different versions of the questionnaire (long and short form). Furthermore, depending on the level of responsibility, information was provided in different level of details, thanks to a drill-down mechanism, in order to explode the data item down to the maximum level of detail represented by municipality or by enumerator.

12. **Buildings** – this area relates to the housing census and includes a registration form and monitoring reports. In order to help enumerators to move around municipalities, the System allowed to download PDF documents with all municipalities addresses divided by tracts.

13. **Questionnaires** – includes all functions strictly connected to the survey; the system provided a form for recording the questionnaires returned from municipalities in order to give a prompt monitoring of this sort of information. Authorized municipalities were able to carry out data entry using the same web application that was made available to households. In order to collect primary variables, required for data dissemination, the system provided a special data entry function available for municipalities that were not authorized to insert data. Any online questionnaire could be printed by operators.

14. **Census Tract** – includes all functions that guide the enumerator's field work. The core of this area is the census tract agenda, which keeps track of changes in the status of each questionnaire. It is initially populated with data from the Population Register and with the under coverage identified in a previous step, and the list of all addresses of each census

tract. The agenda is continuously updated with data from the collection system and input from the enumerators and back office operators.

15. **Census and population register** – it is a dashboard for enabled operators to monitor and manage the differences between the population register and census results. Every questionnaire is processed and operators indicate if individuals included in the questionnaire are the same as reported in the public register or, if necessary, they add the new ones. At the end of this work the system provides four different reports: (a) citizens living in the same dwelling, (b) citizens living in another dwelling, (c) citizens included in the questionnaire but not included in the public register, (d) citizens included in public register but not included in the questionnaire.

16. **Utilities** – includes a set of network support functions spanning the entire survey process.

III. The Online Questionnaire (QPOP)

17. The online Questionnaire of the Italian Population Census (**QPOP**) is a web application made available to citizens that could use it to fill up their census questionnaire in a way which was perfectly equivalent to fill up the printed form. An authorization code was printed in the front page of the paper questionnaire that, together with the respondent's fiscal code, could be used as an authentication credential for accessing the web application. The application reproduces all the three types of paper questionnaires, respectively the two questionnaires for families and households (in long and short version) and the questionnaire for cohabitations.

18. At the end of the data collection phase almost 8,400,000 questionnaires were returned through the QPOP application, corresponding to a percentage of the 33 per cent of the total number of expected questionnaires. The average load in the first two months of Census operations was 100 questionnaires sent per minute, peaking to 300 in periods of maximum load.

19. QPOP was seamlessly integrated with SGR: once a respondent completes a questionnaire in QPOP, the state of the questionnaire is immediately updated in the central database and visible to operators through SGR. Operators could also use QPOP for performing data entry operations.

20. Since the QPOP application was potentially available to the whole Italian population, it was of primary importance to carefully design it so that it could scale gracefully to a huge number of users, at the same time achieving a high robustness (i.e., no application errors presented to the user) and preserving data security despite all the possible security threats. Sophisticated technical solutions were adopted in the internal design of the application for i) reducing the load to the database even in presence of a high number of users connecting to the application and ii) performing data validation at multiple points in the application for guaranteeing consistency of data sent by the user.

21. Besides this, the design largely exploits metadata, avoiding the development of redundant source code for the three questionnaires. Everything that appears in the web user interface (including warning and error messages) is stored in specific metadata tables and resource files. In particular the text of each single question constituting the questionnaire is stored in a (meta-)database table with the 3 different localized (Italian, German and Slovenian) versions. Also the single possible response modalities are stored in a (meta-)database table with the 3 different localized versions. Changes and corrections to the texts constituting the questionnaires could be (and indeed were) made at any moment, even after

the final deployment of the application and its release to citizens, without affecting the source code of the application.

22. Texts are not the only part of the QPOP application stored as metadata. A further fundamental feature of QPOP is that also part of the "behavior" of the application was stored in metadata. This is closely related to the concept of "questionnaire graph", which is a fundamental part of the technique, used to formally model the structure of the questionnaire and the correct set and sequence of questions to be filled in by the respondent.

23. A questionnaire graph is a Directed Acyclic Graph (DAG), such that: (i) nodes are in 1-1 correspondence with each question of the questionnaire; (ii) a (directed) edge from node (question) N_i to node (question) N_j correspond to the fact that the user has to respond to question N_j after having given a response to node N_i. In general, however, there may be more than one edge exiting from the same node N_i (e.g. "If you have answered Yes, then go to question X.Y, otherwise proceed"). As a consequence edges can have labels, i.e. conditions based on question responses. The conditions on the edges exiting from any given node N_i have to be mutually exclusive (and one is necessarily true) and determine which is the actual "next" question after N_i.

24. The questionnaire graph, i.e. the data about nodes, edges (and related conditions) is stored in the QPOP (meta-)database and used by the application (both on client and server side) to visually enable and disable questions on the web page and to validate the user's input before saving the user's responses in the micro data tables. Since the graph is invariant, its structure is "cached" in main memory by the QPOP application, so as to avoid unnecessary and inefficient repeated accesses to the database.

25. Given the questionnaire graph and the responses already given by the user, the application automatically determines the questions to be enabled and disabled in the web user interface. Thanks to the fact the questionnaire graph is actually a DAG, the algorithm to update the enabling/disabling of questions is particularly efficient: in particular it can be shown that each edge has to be considered only once to update the enabling property of the whole questionnaire. In a few words the idea underlying the algorithm is that, given the responses already provided by the user, a question has to be enabled if the user will necessarily reach it, independently of the responses she/he has not given yet. On the contrary it has to be disabled if she/he will certainly not reach it or there is a sequence of responses for the still unanswered questions that will prevent her/him from reaching that question.

26. QPOP also implements a sophisticated mechanism for automatic encoding of textual responses, based on a dictionary, similarity string comparison and automatic ranking. This mechanism was used for the question on highest educational qualification. The corresponding classification dictionary is constituted by more than 6,000 distinct items and it was required that QPOP would provide an already encoded title (in other terms QPOP could not store a free text to be later encoded, either manually or semi-automatically). A selection based on a dropdown list was obviously unfeasible due to the large number of possible items.

27. As a consequence, a mechanism was chosen that closely resembles the interaction provided by search engines: the user enters a list of words describing her/his full title of educational qualification; the system replies by proposing a list of titles taken from the available dictionary, and ranked according to their similarity to what expressed by the user. Finally, the user can select one of the proposed items or try a new search if the proposed list is unsatisfactory.

28. In order to increase the efficiency of the search engine and avoid system servers overloading, a pre-processing of the dictionary is performed. This pre-processing is fairly

heavy from the computational point of view and is consequently performed by an offline specific procedure. The basic steps of the procedures are the following: (i) normalization of the characters in the dictionary items (e.g. accented letters are replaced by the corresponding unaccented ones); (ii) pruning of the "stop words" (e.g. articles, conjunctions and other "useless" words) from the dictionary items; (iii) extraction of the single words from the normalized and pruned versions of the dictionary items produced by the previous steps. These words (called d-words in the following) are stored in the database and linked to the original versions of the dictionary items for later use by the online search engine.

29. The online search engine receives as input (from the user) a generic search string constituted by some words separated by blank spaces. First of all the search string is normalized, pruned and split into a collection of single search words, similarly to what made for the dictionary items. Then the following steps are performed: (i) each search word is compared with the d-words in the database: if the similarity is above a given threshold value, then the d-word is added to the list Lt of words above threshold; (ii) all dictionary items linked to words in the list Lt are extracted from the database; (iii) the items extracted in the previous step are sorted, according to several criteria, particularly: exact or similarity match, frequency of each word in the dictionary, number of matching words, total number of words in a given item.

IV. The network portal (RETE)

30. The Network Portal is a mainly informational web site, organized in two horizontal menu navigation bars which provide access to the various information areas. There are 7 information areas that assist operators in different ways.

31. The **Data Collection Documents** area contains documents which are also working tools to conduct the Census. In addition to the family and housing Questionnaires in PDF format, this area also contains instruction manuals and documents that may help with the compilation of some questionnaire sections (i.e., support measures and codes).

32. The **Instruments** area contains software that helps citizens in choosing the right economical and work activity, while the **Documents** area contains official documentation on connected topics, such as the main reference legislation on Population and on personal data protection.

33. Other areas include **Questions & Answers** (FAQ), **Glossary**, **Video Tutorials** and **Training**, the latter containing training materials for the census network, including an interactive version of the questionnaire, instruction manuals divided into chapters and various slides on the content of classroom training sessions.