A web based management system for addressing census complexity: the Italian experience

Mauro Bruno\(^1\), Maura Giacummo\(^2\), Marco Silipo\(^3\), Giulia Vaste\(^4\)

\(^1\) Italian Statistical Institute (ISTAT), e-mail: mbruno@istat.it
\(^2\) Italian Statistical Institute (ISTAT), e-mail: magiacum@istat.it
\(^3\) Italian Statistical Institute (ISTAT), e-mail: silipo@istat.it
\(^4\) Italian Statistical Institute (ISTAT), e-mail: vaste@istat.it

Abstract

The Population Census is inherently a complex survey; further the last Italian Census introduced important methodological innovations such as a transition towards a register-supported census and several ways for respondents to return questionnaires. Such innovations increase the possible errors and the survey complexity. Therefore the problem to be addressed concerns how to perform a *quality* census, which is innovative in methodology as well as effective both in costs and in data results. The Italian Statistical Institute (Istat) answer to these issues is a census focused on web technologies. This paper shows how an integrated web information system, based on standard software frameworks and common best-practices, has been used to manage census workflow that involves multiple phases, integrates different data sources and is used by several kinds of actors. The success of the Population Census experience leads Istat to adopt such technological infrastructure as the core of the ‘Continuous Population Census’.

**Keywords**: quality census, multiple data sources, standard framework.

1. Introduction

The 2011 Italian Population Census introduced important innovations compared with the traditional census. The main innovations are: a transition towards a register-supported census; questionnaire personalization with information concerning the family holder and delivery of the questionnaires by the Italian postal service; several ways for respondents to return their questionnaires, such as online compilation, delivery to postal office or to census bureau or to enumerators (Crescenzi and Mancini (2010), UNECE (2012)). As a result, the census methodological and procedural process becomes more effective and efficient, but also more complex and difficult to manage.

Moreover, like all statistical surveys, the census survey is inherently affected by errors, especially when it is register-supported. For example, here are some issues that may arise: under-coverage, over-coverage, error-prone lists, missing questionnaire information, and duplicate questionnaires. Such issues are due both to the use of lists and to respondents’ behaviour.

Thus, we have to provide an answer to the following questions: how can we minimize possible errors? And how can we perform a *quality* census, which is (i) innovative in methodology as well as effective both in costs and in data results; (ii) easy to be managed by the census operators; and (iii) that can be monitored in all its complex phases?
The Italian Statistical Institute’s (Istat) answer to these issues is a census focused on web technologies. Istat developed an integrated web information system that supports all the different phases of the survey process. Such system is composed of three web applications: a) the online questionnaire (QPOP) used both by the citizens to provide their census data, and by the enumerators to perform the data entry of the paper questionnaires collected; b) an online documentation system publishing different kinds of contents such as wiki, faq, manuals, legislative references, etc; c) a web based management and monitoring system (SGR) accessible to the census operators, supporting the survey process activities.

QPOP guides citizens in the correct compilation of their questionnaire through consistency rules and error checking. Moreover QPOP presents to the respondent only the correct set and sequence of questions to be filled in, so that the online compilation turns out to be easier, faster and less error prone than the printed form. As a result the quality of the collected data is increased with respect to the traditional census. A significant part of Italian citizens chose QPOP for questionnaire compilation, and about 8.500.000 questionnaires, i.e. more than 33% of the total number of expected questionnaires, were returned. The broad use of QPOP resulted in a considerable reduction of the enumerators’ field work with significant organizational and cost-effective advantages. Further, the quality of the collected data allowed a quick production of the first census results (Virgillo A., Tininini L. (2012)).

The online documentation system acts as the reference site for the network of SGR operators, who use it to access documentation material, such as manuals and legislative references, but also as an up-to-date information site for contents such as wiki, faq and communications.

SGR is a complete instrument that helps census operators during all the survey phases. It replaces previous monitoring systems, which offered a dashboard with quantitative indicators for tracking questionnaires, without providing any management or support tools to census operators. SGR has been designed to provide to the different users of the system: (i) up-to-date information at different aggregation levels, including at single questionnaire level; (ii) a tool for cooperative working, guided through a forced workflow of questionnaire life-cycle. Such system helps reduce the possible statistical errors by checking and controlling potential register errors; by monitoring “in itinere” the survey; and by managing enumerators’ field work.

The rest of the paper is organized as follows: in Section 2 the system architecture is presented, focusing on the technological innovations introduced in the last Italian population census; in Section 3 the Italian experience is illustrated and the main functionalities of SGR are shown, pointing out how SGR supports the complex census process, allowing cooperative operators work and data integration; in Section 4 the concluding remarks are discussed.

2. System architecture

One of the main complexities of setting up such an important and business critical system like SGR is the integration of multiple sources of data and the ability to guarantee the availability of the whole system on a 24/7 basis, in particular during peak hours. These goals can be achieved only by designing and implementing a flexible architecture, which is based on balanced and replicated systems and uses a consolidated middleware infrastructure where components can be plugged easily and without massive code reworks.
As shown in Figure 1, SGR is a core part of such architecture, integrating its modules with the questionnaire application that was used by respondent citizens for data entry but also from SGR operators entering data from the paper questionnaires. Data on which SGR is based relies on a very complex ORACLE database which has to be accurately tuned for data access optimization from a very high amount of users, including citizens and operators.

The core software infrastructure of SGR and QPOP follows the Model-View-Controller design pattern. The main technology on which SGR is based is the Java2 Enterprise Edition (J2EE) platform. Java was the core language used for the development of the last Agriculture Census management system. The team that was in charge of the development of SGR for the Population census took advantage of the previous experience, improving the software in terms of security and general quality, reusing many parts of the infrastructure but switching to the state-of-the-art versions of the frameworks chosen as building blocks. The heavy use of frameworks has proven to provide a significantly positive impact, producing a cleaner code that was easy to write, test and maintain, resulting in a more robust application. The implementation exploits in particular three widespread open source frameworks, in particular: Struts2, Spring and Hibernate. Hibernate proved to be fundamental for developers, as it acts as a software layer which eases working with database tables through a simple model based on Java objects, called “beans”. This speeds up development time and delegates the SQL interaction with the DBMS to the framework: programmers only have to deal with the generally more familiar Java syntax.

Another technology widely used in SGR is AJAX: this Javascript-based technology acts as glue between the GUI and the server-side components of the system, allowing a tight interaction among the controls available on the end-user interface (buttons, lists, collapsible data sections) and consequent real-time actions happening on the server. This technology brings real improvements to the end-user, who is actually guided in his work since the data SGR presents to him is always fetched from the database using small and focused queries.
This also brings important performance benefits: for example, when the user selects a district from a dropdown list and is called to select a city to work with, the cities he can choose from are a subset which depends on the previous choice of the district.

Another component of the architecture is the online documentation portal that has been developed by means of a widely popular open source CMS (Typo3), which is written entirely in PHP. This led to an additional integration problem due to the fact that Typo3’s database management system is MySQL while SGR data is Oracle-based: since any SGR user had to be authenticated on both systems, user data had to be constantly synchronized, through a custom script.

The components of the system architecture are shown in Figure 1. QPOP, SGR and the Documentation system are three integrated web applications. The integration mainly concerns the authentication system and the management of questionnaires returns. With respect to the authentication system, on the one hand, a single-sign-on mechanism is provided for SGR and QPOP, allowing an SGR operator to access QPOP for the data entry of his assigned questionnaires. On the other hand, SGR and the Documentation system authentication tables are synchronized allowing a user to access both applications with the same credentials. Concerning the questionnaire returns, QPOP and SGR share information on the questionnaire status: when a citizen completes the online compilation of his questionnaire this information is available in SGR; when a questionnaire is returned and registered in SGR the online compilation is disabled.

The database is loaded with normalized data from different sources, e.g. House numbers enumeration database, LAC, Revenue agency, Postal service. A detailed description of the integration of such sources is provided in Section 3.

3 The Italian census experience

SGR has been designed as a collection of functions, each related to a process phase and customized according to the user profile that accesses it. About eighty functions, grouped in menus and sub-menus, have been developed to support all the census activities (as shown in Figure 2). Therefore SGR is a modular, flexible, and scalable system which allows an agile development process. Although more than twenty developers were involved in the realization of SGR, the plug-and-play design of SGR allowed a strong cooperation and a rapid development of all the functionalities.

The system is dynamically customized according to the profile of the logged user: both the menus and sub-menus are personalized displaying only the functions the operator is authorized to use.

To illustrate the functionalities offered by SGR we focus on the following macro-areas: a) integration of multiple data sources (registers, civic number enumeration survey, etc.); b) interaction of different actors of the survey process (enumerators, supervisors, Istat, postal service); c) implementation of the workflow for the questionnaire life-cycle management; and d) up-to-date monitoring of the survey progress.
3.1 Integration of multiple data sources

An important innovation of the 2011 Italian Population Census has been the use of local population register (LAC): questionnaires have been personalized with information concerning the family holder and delivered by the Italian postal service. In particular, the SGR database has been loaded with normalized data, almost sixty millions of individuals, coming from LAC dating to 31/12/2010. Further, to take into account population flows in the period between 31/12/2010 and the census date, i.e. 9/10/2011, a second data loading was performed for cities with more than twenty thousand citizens. Small cities have been supported in such operations by SGR through suitable functionalities.

LAC can be affected by errors due to the incompleteness of the register. Moreover, taking into account only one data source results in missing out individuals not present in the register but living in the municipality (under-coverage). So, data provided by different sources, such as revenue agency, permits to stay and civic number enumeration survey, have been integrated in the system to help detect under-coverage. In particular, civic number enumeration survey and municipal address archives allowed the detection of buildings with no corresponding LAC individual: almost nine millions possible under-coverage signals have been identified in the system and checked by the enumerators.

SGR manages also the different questionnaire returns offered to respondents, i.e. web compilation, returns to postal office and to enumerators. On the one hand the integration between QPOP and SGR allows a real-time monitoring of the online compilations; on the other hand the integration with services provided by the Italian postal service is necessary in order to load information about both deliveries of the personalized questionnaires to the respondents and returns to the postal offices.

3.2 Interaction of different actors of the survey process

In order to take into account Italian different geographical, socio-demographical and housing situations, the survey has to be carried out at a fine-grained level, i.e. municipal sub-areas (census areas). Thus SGR includes several user profiles each characterized by a territorial visibility (national, regional, provincial, municipal) and available functions (Table 1).

User profiles with national visibility have monitoring functionalities that enable one to monitor the survey progress on the whole territory and to take strategic decisions during the survey. User profiles with regional visibility have been provided to Istat regional employees. The regional employees’ duty is to organise and coordinate the survey on their territory and to support the municipal operators. The municipal operator is the key profile in SGR: it is in fact responsible for
all operative phases of the census. Such users can: (i) define the local survey network, i.e. supervisors and related enumerators; (ii) assign census areas with related questionnaires to the enumerators and (iii) monitor the progress in the questionnaires life-cycle. Enumerators and supervisors carry out field work and back-office activities, such as registration of the questionnaire returns to the census bureau, interviews to the respondents, check of data provided by the civic number enumeration survey in order to detect under-coverage, etc.

Actually, SGR gave the possibility to create autonomously the survey network, inserting more than eighty five thousand operators. This is a significant result since the creation of the survey network is not managed as a centralized task but is distributed on the whole territory.

<table>
<thead>
<tr>
<th>User profile</th>
<th>Functionalities</th>
<th>Territorial visibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Istat</td>
<td>Monitoring functionalities at all territorial level, i.e. national, regional, provincial, municipal and census areas level</td>
<td>National</td>
</tr>
<tr>
<td>URC</td>
<td>Monitoring and support functionalities at regional, provincial, municipal, and census areas level</td>
<td>Regional</td>
</tr>
<tr>
<td>UPC</td>
<td>Monitoring and support functionalities at provincial, municipal, and census areas level.</td>
<td>Provincial</td>
</tr>
<tr>
<td>UCC</td>
<td>Monitoring and support functionalities at municipal and census areas level. Functionalities for creating the municipal operators network. Functionalities for assigning, coordinating and supervising both enumerators and supervisors work. Operative functionalities for field work and back-office activities.</td>
<td>Municipal</td>
</tr>
<tr>
<td>CoC</td>
<td>Functionalities for assigning, coordinating and supervising enumerators work. Operative functionalities for field work and back-office activities.</td>
<td>Assigned census areas</td>
</tr>
<tr>
<td>Ril</td>
<td>Operative functionalities for field work and back-office activities</td>
<td>Assigned census areas</td>
</tr>
</tbody>
</table>

### 3.3 Implementation of the workflow for the questionnaire life-cycle management

SGR guides the operators to fill out the survey correctly, offering a fixed path through the questionnaire working phases. Each questionnaire working phase has been linked to a state. The transition between two different states is realized through SGR functions or through external operations, such as the completion of the online compilation. Each function has both pre-condition states, i.e. states that allow the function use, and post-condition states, i.e. states assigned by the function to the questionnaire. In this way SGR defines a predefined flow of questionnaire states, which guides and forces the questionnaire life-cycle. Such questionnaire life-cycle management allows also cooperative operators work. For example we can consider the following scenario: 1) a back-office operator registers the paper questionnaire return; 2) SGR updates the questionnaire state; 3) consequently the enumerator, responsible for that questionnaire management, is informed of the return and can proceed with following working phases.

Thus SGR is a distributed workflow system, in which, on the one hand, each operator works autonomously and, on the other, a centralized monitoring of the overall census activities is provided. As a result, using SGR as a survey tool allows for cost-effectiveness, real-time management, support for cooperative work and on-going monitoring.
The most important function in SGR is the “diary”: a control panel that shows to the survey operator an up-to-date list of his assigned questionnaires, built by the different sources described in section 3.1. Each element of the list displays the respondent’s personal data, the questionnaire state, and the operations performed on the questionnaire. The operator can modify the questionnaire state: on the basis of the current questionnaire state the diary shows the possible next states (including the final states, i.e. the survey results). Further the diary displays the information to be edited according to the state transition. In such a way the diary allows one to manage the significant aspects of the process, such as the coexistence of different questionnaire returns offered to respondents. Each return sets the questionnaire in a defined state, through internal or external functions, i.e. web compilation, loading of postal office information in the system database and census operator registration. Such information, which is visible in the diary, allows the complete monitoring of the returns. As a result, the enumerators are now able to interview only the respondents that did not return the questionnaire. The enumerator’s work is thus more efficient, and the quality of survey data is increased due to a reduction of potential errors by enumerators.

Through the diary is also possible to edit information concerning primary variables such as the number of males, females, and foreigners, etc. Such information is mandatory in order to reach a questionnaire final state. The availability of primary variables data in SGR has been a key element for a rapid dissemination of provisional data.

The diary results a complete instrument for all the operators involved in the survey process: the enumerators use it as a control panel that shows them an up-to-date list of assigned questionnaires and guides their field work; the supervisors use it as instrument to monitor the work of their assigned enumerators; the municipal operators refer to the diary to monitor the survey progress in the census areas.

### 3.4 Up-to-date monitoring of the survey progress

SGR permits a comprehensive monitoring of the survey process. On the one hand, SGR offers a detailed supervising of all the operations performed on each questionnaire, displaying author and date of each operation. On the other hand, several reports are provided in order to show the progress of each survey phase: delivery of the questionnaires by the postal service; returns of the questionnaires; survey progress on the basis of questionnaire states; activity of the operators (with a high level of detail).

Reports are available at different data aggregation level. In particular, it is possible to display data concerning a particular operator or data at different territorial levels, i.e. national, regional, provincial, and municipal.

All these reports are dynamically generated using snapshots of data taken at regular intervals, producing tabular views of reported data. Every report can be exported in the most common formats like XLS and PDF, but also in a CSV plain text file, allowing users to import data in any custom tool.
4. Concluding remarks

SGR turned out to be a very useful instrument that supported the 2011 Italian Population Census: it has been daily accessed by more than one hundred thousand users during the survey; it supported a complex survey with about 8,500,000 online questionnaires compilation (33% of the total returns), 7,826,000 census bureau returns, over three million questionnaires returned to the enumerators and the remaining 20% returned to postal offices.

Even though measurements on the quality of the survey results have not been performed yet, the management of the survey pointed out significant improvements towards a cost-effective and quality census. SGR greatly contributed to achieving such results.

First, it supported a register-based census, allowing a reduction in the related errors. Such result has been achieved through the use of multiple data sources as well as a constant and complete monitoring of all survey phases.

Second, SGR offered a predefined and forced workflow in the questionnaire life-cycle management; this enables enumerators’ work to be less prone to errors and more cooperative.

Third, SGR has been used by operators with different responsibilities, providing suitable functionalities for each profile: (i) the diary effectively supported the enumerators’ field work and (ii) reports allowed monitoring to users with different territorial visibility, also supporting the strategic decisions of Istat census managers.

SGR has been originally adopted in the 2010 Agricultural Census, and it is actually in use in the 2011 Industry and Services Census survey. In this survey, SGR has been enriched with new functionalities, such as micro data check on the base of predefined rules. Due to SGR the data production process is agile, efficient and high quality.

Since SGR software architecture is mainly framework-based and standard compliant, it has demonstrated to be very suitable to constitute the foundation of a set of additional web systems that have been recently implemented in order to support the subsequent stages of the Census process. As few but significant examples worth mentioning are: the Post Enumeration Survey (PES), the System for the Review of the Civil Registries (Sirea), and the Survey on Woody Plantations, etc.

Istat technological innovation plans include an SGR generalization with the aim to adopt it as a system for the management of different surveys. In particular, it will be a central component in the design of the ‘Continuous Population Census’, which is a main topic in Istat’s strategic plan.

References

