



Conseil économique et social

Distr. générale
14 mars 2012
Français
Original: anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Groupe d'experts des recensements de la population et des habitations

Quatorzième réunion

Genève, 24 et 25 mai 2012

Point 5 de l'ordre du jour provisoire

Collecte de données par Internet

Système d'information en ligne utilisé pour le recensement de la population italienne

Note de l'Institut national de statistique italien (Istat)

Résumé

Le quinzième recensement de la population italienne a introduit plusieurs innovations pour l'appui aux nombreuses activités du processus de collecte, notamment l'utilisation et l'analyse du registre de population local (*Anagrafe*), la participation des services municipaux locaux au processus de recensement tout entier et une stratégie de collecte de données par divers canaux. Le présent document décrit les principales caractéristiques du système en ligne utilisé.

I. Introduction

1. Le présent document a été élaboré par Maura Giacummo, Leonardo Tininini, Marina Venturi et Antonino Virgillito de l'Institut national de statistique italien (Istat).
2. Le quinzième recensement de la population italienne a introduit plusieurs innovations pour l'appui aux nombreuses activités du processus de collecte. La première concernait la stratégie de recensement fondée sur l'utilisation et l'analyse du registre de population local (*Anagrafe*), qui a été utilisé comme base d'informations pour le recensement. En outre, une comparaison a été effectuée entre les données sur les citoyens recueillies dans le cadre du recensement et celles inscrites dans le registre, et le réalignement correspondant a été opéré. Une deuxième innovation a consisté à assurer la participation des services municipaux locaux à l'ensemble du processus de recensement, notamment à chaque étape de l'enquête, depuis la collecte des données jusqu'à leur contrôle et à l'enregistrement des principales variables en vue de la diffusion de données provisoires et dans certains cas également du questionnaire tout entier. La troisième et peut-être la plus importante innovation visait les techniques de collecte de données par divers canaux. De fait, la collecte de données classique sur support papier effectuée par des agents de recensement dans le cadre d'entretiens en face à face a été complétée par un questionnaire électronique qui peut être utilisé à la fois sur Internet par les citoyens, lesquels procèdent directement à la compilation des données, et par le personnel du réseau qui enregistre les données précédemment recueillies sur des documents papier. Le présent document décrit les principales caractéristiques du système en ligne, constitué de trois applications intégrées dans une plate-forme unique: le système de gestion des recensements de la population (SGR), le questionnaire en ligne (QPOP) et la documentation en ligne destinée aux opérateurs (RETE).
3. Le système de gestion des recensements de la population (SGR) est accessible aux opérateurs des bureaux de recensement et au personnel d'Istat. Ses fonctions couvrent les différentes étapes de la collecte de données. Les opérateurs peuvent notamment gérer le processus en assignant des ménages aux agents de recensement, en contrôlant la collecte et en surveillant les données. Comme la collecte se fait par divers canaux, le système permet de contrôler les questionnaires collectés par les différents moyens possibles (en ligne, par les recenseurs ou par envoi aux centres de collecte postaux/municipaux). Les statisticiens y ont également recours pour recueillir les principales variables à diffuser dans les résultats provisoires du recensement. Enfin, grâce à un autre module fondamental du système, les services administratifs locaux peuvent comparer les données sur les citoyens collectées dans le cadre du recensement à celles qui sont inscrites dans les registres de population locaux et à les réaligner en conséquence.
4. Le questionnaire en ligne permet aux citoyens italiens de remplir leur questionnaire en ligne. L'interface aide les utilisateurs à se conformer aux bonnes règles de compilation et à vérifier la présence d'erreurs avant de présenter le questionnaire final. Une fois complété, celui-ci est immédiatement accessible dans le SGR. Le questionnaire en ligne peut aussi être utilisé par les opérateurs des centres de recensement pour la saisie de données. Le choix d'une seule plate-forme en ligne comprenant à la fois le questionnaire en ligne et le système de gestion était avantageux pour plusieurs raisons: contrôle en temps réel du processus, amélioration de la qualité des données, réduction des activités sur le terrain, réduction de l'écart entre la collecte et le traitement des données. Ainsi, une coopération efficace s'est instaurée entre Istat et le réseau de recensement de la population. Dans ce cadre, 100 000 opérateurs environ ont participé aux activités de recensement. Plus de 30 % des citoyens ont rempli leur questionnaire en ligne, témoignant ainsi du succès particulier de ce moyen.

II. Système de gestion des recensements de la population (SGR)

5. Un système fondé sur les technologies de l'information a été mis en œuvre pour aider le réseau de recensement à gérer les différentes étapes du recensement de la population et des habitations. Le système de gestion et de contrôle permet aux services d'appui municipaux de suivre les questionnaires reçus de divers canaux et d'orienter le travail des agents de recensement sur le terrain, à savoir la collecte des questionnaires manquants et le traitement des taux de couverture insuffisants ou excédant les 100 %.

6. Les innovations décrites dans l'introduction devaient être complétées par certaines fonctionnalités qui se sont ajoutées au système de contrôle ordinaire: synchronisation et échange de données avec les systèmes externes; synchronisation avec le système de collecte de données en ligne; complexité accrue des fonctions de contrôle due à la diversification des questionnaires et à la multiplicité des modes de renvoi des questionnaires aux communes; utilisation des informations pour orienter l'activité quotidienne et systématique des agents recenseurs. Pour cette raison, une application spécialisée fondée sur l'utilisation des technologies du Web a été développée, à savoir le système de gestion des enquêtes (*Sistema di Gestione della Rilevazione*) ou SGR. Elle assure le maximum de sécurité pendant les phases de transmission et de stockage des données, conformément aux règles types de l'Institut national de statistique.

7. Le système de gestion peut être considéré comme un système de flux de travail réparti dans lequel chaque opérateur peut travailler de façon indépendante suivant une procédure clairement définie. De plus, ce modèle permettait de gérer les processus de production (suppression des questionnaires, changement d'état, etc.), Istat pouvant ainsi éviter des problèmes qui n'auraient pu être résolus que par une intervention manuelle sur des questionnaires isolés.

8. Cette procédure opératoire a été utile en termes d'actualité, de qualité des données et de coûts. Le système a pu gérer une enquête de recensement portant sur plus de 25 millions de ménages, soit 60 millions d'individus. Par ailleurs, il a offert un accès simultané à plus de 100 000 opérateurs (réseau de recensement) travaillant chaque jour et sept heures par jour. Il comporte plus de 70 fonctions regroupées par type et organisées en 8 macrodomaines.

9. **Changements intervenus dans le registre de population local** – Ce domaine est consacré aux petites municipalités (de moins de 20 000 habitants), qui peuvent introduire dans le système des changements intervenus dans les listes du registre de population local envoyées à l'Istat entre le 30 décembre 2010 et le 9 octobre 2011 (date du recensement). Des formulaires spéciaux permettent de modifier les informations (nouveaux ménages, familles ayant emménagé dans les municipalités ou en étant parties). D'autres municipalités ont envoyé leur registre de population local à l'aide d'une application Web spéciale et les données correspondantes ont été enregistrées dans le SGR.

10. **Opérateurs** – Ce groupe de fonctions permet aux municipalités de gérer de façon autonome le personnel de terrain et d'appui. Les utilisateurs autorisés peuvent saisir, visualiser et modifier des informations personnelles des coordonnateurs et des recenseurs. Chaque membre du réseau qui a utilisé le système a été enregistré par ses superviseurs au moyen de formulaires spéciaux. Cette opération était essentielle pour la création du profil et du mot de passe utilisateur qui permettaient d'accéder au système. Il a été attribué à chaque utilisateur un profil qui l'autorisait à voir uniquement les fonctions relevant de ses attributions. Pour des raisons de sécurité, le mot de passe a été envoyé à l'adresse électronique saisie lors de l'enregistrement et devait être modifié lors du premier accès au système. Le nouveau mot de passe est enregistré dans le système après cryptage. Les opérateurs autorisés ont également la possibilité de rattacher des agents recenseurs à des

coordonneurs, d'assigner des secteurs de recensements (et des questionnaires) aux agents recenseurs.

11. **Rapports récapitulatifs** – Il s'agit d'un ensemble de rapports qui permet de contrôler en continu l'évolution de l'enquête et le travail des recenseurs. Ces rapports permettent une évaluation «presque en temps réel» de l'état d'avancement des opérations sur le terrain. Le contrôle devait intégrer divers types de données liés aux différents modes de diffusion du questionnaire et aux différentes versions (longue ou courte) de celui-ci. De plus, selon le niveau de responsabilité, les informations étaient fournies à des niveaux de détail différents, au moyen d'un mécanisme de progression descendante permettant d'éclater les données à un niveau maximum de détail représenté par la municipalité ou le recenseur.

12. **Bâtiments** – Ce domaine se rapporte au recensement des habitations et comprend un formulaire d'enregistrement, ainsi que des rapports de contrôle. Pour aider les agents de recensement à se déplacer dans les municipalités, le système a autorisé le téléchargement de documents PDF, toutes les adresses des municipalités étant subdivisées en secteurs.

13. **Questionnaires** – Ce domaine comprend toutes les fonctions strictement liées à l'enquête; le système a prévu un formulaire pour l'enregistrement des questionnaires renvoyés par les municipalités afin d'accélérer le contrôle de ce type d'information. Les municipalités autorisées ont pu saisir des données à l'aide de la même application Web que celle mise à la disposition des ménages. Pour recueillir les principales variables nécessaires à la diffusion des données, le système a mis une fonction de saisie de données spéciale à la disposition des municipalités non autorisées à introduire des données. Tout questionnaire en ligne pouvait être imprimé par les opérateurs.

14. **Secteur de recensement** – Ce domaine comprend toutes les fonctions qui orientent les recenseurs dans leur travail sur le terrain. Son élément central est l'agenda du secteur de recensement, qui garde des traces des changements d'état de chaque questionnaire. À l'origine, il est alimenté par les données issues du registre de la population, compte tenu du taux de couverture insuffisant identifié dans une étape précédente, et par la liste de toutes les adresses de chaque secteur de recensement. L'agenda est actualisé en permanence au moyen des données provenant du système de collecte ou saisies par les agents de recensement et les opérateurs des services d'appui.

15. **Recensement et registre de population** – Il s'agit d'un tableau de bord qui permet aux opérateurs autorisés de contrôler et de gérer les différences entre le registre de la population et les résultats du recensement. Chaque questionnaire est traité et les opérateurs indiquent si les individus qui y sont inclus sont les mêmes que ceux qui sont inscrits dans le registre public ou, s'il y a lieu, ajoutent les nouveaux individus. À la fin de ce travail, le système établit quatre rapports différents portant sur les sujets suivants: a) citoyens vivant dans le même logement; b) citoyens vivant dans un autre logement; c) citoyens inclus dans le questionnaire mais pas dans le registre public; d) citoyens inclus dans le registre public mais pas dans le questionnaire.

16. **Utilitaires** – Ce domaine comprend un ensemble de fonctions d'appui au réseau visant le processus d'enquête tout entier.

III. Questionnaire en ligne (QPOP)

17. Le questionnaire en ligne (QPOP) utilisé dans le cadre du recensement de la population italienne est une application Web accessible aux citoyens qui pouvaient l'utiliser pour remplir le questionnaire de recensement exactement comme s'ils complétaient un formulaire imprimé. Un code d'autorisation a été imprimé sur la première page du

questionnaire papier, ce code pouvant être utilisé avec le numéro d'immatriculation fiscal du répondant pour authentifier l'accès à l'application Web. Celle-ci reproduit les trois types de questionnaires papier, respectivement deux pour les familles et les ménages (en versions longue et courte) et un pour les cohabitations.

18. À la fin de l'étape de collecte de données, près de 8 400 000 questionnaires ont été renvoyés au moyen de l'application QPOP, soit 33 % du nombre total des questionnaires attendus. Au cours des deux premiers mois du recensement, la charge moyenne était de 100 questionnaires envoyés par minute, avec un pic de 300 en période de charge maximale.

19. Le questionnaire en ligne a été intégré en douceur dans le SGR: lorsqu'un répondant a rempli ce questionnaire, l'état de celui-ci était immédiatement actualisé dans la base de données centrale et pouvait être visualisé par les opérateurs grâce au SGR. Les opérateurs pouvaient aussi se servir du questionnaire pour saisir des données.

20. L'application QPOP étant potentiellement accessible à l'ensemble de la population italienne, il était primordial de la concevoir avec soin pour qu'elle puisse aisément être étendue à de nombreux utilisateurs tout en présentant une grande robustesse (c'est-à-dire que l'utilisateur ne devait être confronté à aucune erreur) et en préservant la sécurité des données malgré toutes les menaces possibles. Des solutions techniques de pointe ont été adoptées dans la conception interne de l'application, afin i) de réduire la charge pesant sur la base de données même lorsque les utilisateurs qui se connectent sont nombreux, et ii) de valider les données en de multiples points de l'application pour garantir la cohérence des données envoyées par l'utilisateur.

21. En outre, l'application a été conçue de manière à exploiter largement les métadonnées, ce qui permet d'éviter le développement d'un code source redondant pour les trois questionnaires. Tout ce qui apparaît sur l'interface utilisateur du Web (notamment les messages d'avertissement et d'erreur) est stocké dans des tables de métadonnées et des fichiers de ressources spéciaux. En particulier, le texte de chaque question constituant le questionnaire est mémorisé dans une base de (méta)données avec les trois différentes versions localisées (italienne, allemande et slovène). De même, les seules modalités de réponse possibles sont mémorisées dans une table de la base de (méta)données avec les trois différentes versions localisées. À tout moment, des modifications et corrections peuvent être apportées (et ont effectivement été apportées) aux textes constituant les questionnaires, même après la mise en place finale de l'application et sa mise à la disposition des citoyens, sans affecter le code source de l'application.

22. Les textes ne constituent pas l'élément unique de l'application QPOP stocké en tant que métadonnées. Une autre caractéristique fondamentale est qu'une partie aussi du «comportement» de l'application a été mémorisée dans les métadonnées. Tout cela est étroitement lié au concept de «graphe du questionnaire», élément fondamental de la technique, qui sert à modéliser formellement la structure du questionnaire ainsi que le bon ensemble et la bonne séquence de questions auxquelles le recensé doit répondre.

23. Le graphe du questionnaire est un graphe orienté acyclique (DAG) tel que: i) les nœuds présentent une correspondance biunivoque avec chaque question du questionnaire; ii) une arête (orientée) allant du nœud (question) N_i au nœud (question) N_j représente l'obligation pour l'utilisateur de répondre à la question N_j après avoir répondu au nœud N_i . En général, plusieurs arêtes peuvent cependant sortir du même nœud N_i (par exemple, «Si vous avez répondu par «oui», allez à la question X.Y, sinon continuez.»). Les arêtes peuvent donc être assorties de labels, c'est-à-dire de conditions fondées sur les réponses aux questions. Les conditions relatives aux arêtes sortant de tout nœud N_i donné doivent s'exclure mutuellement (l'une d'entre elles étant nécessairement vraie) et déterminer quelle est effectivement la «prochaine» question après le nœud N_i .

24. Le graphe, c'est-à-dire les données concernant les nœuds, les arêtes (et les conditions correspondantes), est mémorisé dans la base de (méta)données QPOP et utilisé par l'application (du côté client comme du côté serveur) pour valider ou invalider visuellement les questions sur la page Web et pour valider la saisie faite par l'utilisateur avant de sauvegarder ses réponses dans les tables de microdonnées. Le graphe ne pouvant pas varier, sa structure est placée dans le «cache» de la mémoire principale par l'application QPOP afin d'éviter des accès répétés inutiles et inefficaces à la base de données.

25. Compte tenu du graphe et des réponses déjà données par l'utilisateur, l'application détermine automatiquement les questions qui doivent être validées ou invalidées sur l'interface utilisateur du Web. Le graphe étant de fait un graphe orienté acyclique, l'algorithme visant à actualiser la validation ou l'invalidation des questions se révèle particulièrement efficace. Il peut être démontré en particulier que chaque arête ne doit être prise en compte qu'une seule fois pour actualiser la propriété de validation de l'ensemble du questionnaire. En bref, l'idée sous-tendant l'algorithme est que, au vu des réponses déjà fournies par l'utilisateur, une question doit être validée si l'utilisateur doit forcément y arriver, indépendamment des réponses qu'il n'a pas encore données. Par contre, elle doit être invalidée s'il est certain que l'utilisateur n'y arrivera pas ou s'il y a une séquence de réponses pour les questions encore sans réponse qui empêchera l'utilisateur d'arriver à la question considérée.

26. Le questionnaire en ligne met également en œuvre un mécanisme perfectionné de codage automatique des réponses textuelles, fondé sur un dictionnaire, un calcul de la similitude par comparaison de chaînes et un classement automatique. Ce mécanisme a été utilisé pour la question ayant trait aux diplômes les plus élevés. Le dictionnaire de classement correspondant comprend plus de 6 000 rubriques distinctes et il fallait que le questionnaire en ligne fournisse un titre déjà codé (autrement dit, il ne pouvait pas stocker un texte libre à coder ultérieurement, de manière manuelle ou semi-automatique). À l'évidence, une sélection ne pouvait pas être opérée à partir d'une liste déroulante en raison du grand nombre de rubriques possibles.

27. Le choix s'est donc porté sur un mécanisme très semblable aux moteurs de recherche en matière d'interaction: l'utilisateur saisit une liste de mots indiquant l'intitulé complet de son diplôme; le système répond en proposant une liste de titres extraits du dictionnaire disponible et classés selon leur similitude avec ce qui a été indiqué par l'utilisateur. Enfin, ce dernier peut choisir l'une des propositions ou tenter une nouvelle recherche si la liste proposée n'est pas satisfaisante.

28. Pour renforcer l'efficacité du moteur de recherche et éviter une surcharge des serveurs, le dictionnaire fait l'objet d'un prétraitement. Cette opération est assez lourde du point de vue informatique et se déroule par conséquent selon une procédure hors ligne particulière, dont les principales étapes sont les suivantes: i) normalisation des caractères des rubriques du dictionnaire (par exemple, les lettres accentuées sont remplacées par les lettres non accentuées correspondantes); ii) élagage des «mots vides» (par exemple les articles, conjonctions et autres termes «inutiles» des rubriques du dictionnaire; iii) extraction des mots uniques des versions normalisées et élaguées des rubriques du dictionnaire obtenues lors des étapes précédentes. Ces mots (dénommés ultérieurement mots issus du prétraitement) sont mémorisés dans la base de données et rattachés aux versions originales des rubriques du dictionnaire pour être ultérieurement utilisés par le moteur de recherche en ligne.

29. Le moteur de recherche en ligne reçoit en entrée (de l'utilisateur) une chaîne de recherche générique composée de certains mots séparés par des espaces. Tout d'abord, la chaîne de recherche est normalisée, élaguée et subdivisée en un ensemble de mots de recherche, de manière similaire à ce qui se fait pour les rubriques de dictionnaire. Les étapes ci-après se déroulent ensuite: i) chaque mot de recherche est comparé aux mots issus

du prétraitement figurant dans la base de données: si la similitude est supérieure à un seuil donné, le mot issu du prétraitement est ajouté à la liste Lt des mots se trouvant au-dessus du seuil; ii) toutes les rubriques du dictionnaire liées aux mots de la liste Lt sont extraites de la base de données; iii) les rubriques extraites au cours de l'étape précédente sont triées en fonction de plusieurs critères, notamment: concordance exacte ou similitude, fréquence de chaque mot dans le dictionnaire, nombre de mots concordants, nombre total de mots d'une rubrique donnée.

IV. Portail du réseau (RETE)

30. Le portail du réseau est essentiellement un site d'information sur le Web comprenant deux barres de navigation/menu horizontales qui donnent accès aux divers domaines d'information. Ceux-ci, au nombre de sept, offrent une assistance aux opérateurs de différentes façons.

31. Le domaine **Documents pour la collecte de données** contient des documents qui sont également des outils de travail pour le recensement. Outre des questionnaires sur les familles et les habitations en format PDF, ce domaine contient aussi des manuels d'instruction et des documents qui peuvent faciliter la compilation de certaines parties du questionnaire (codes et mesures d'appui).

32. Le domaine **Instruments** contient des logiciels qui aident les citoyens à choisir la bonne activité professionnelle ou économique tandis que le domaine **Documents** contient des documents officiels concernant les thèmes connexes, par exemple les principales lois de référence relatives à la population et à la protection des données personnelles.

33. Les autres domaines sont les suivants: **Questions et réponses** (FAQ), **Glossaire**, **Vidéos pédagogiques** et **Formation**, le dernier domaine contenant du matériel de formation destiné au réseau de recensement, notamment une version interactive du questionnaire, des manuels d'instruction divisés en chapitres et divers diaporamas sur le contenu des sessions de cours formels.
