



Экономический и Социальный Совет

Distr.: General
14 March 2012
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Группа экспертов по переписям населения и жилищного фонда

Четырнадцатое совещание

Женева, 24–25 мая 2012 года

Пункт 5 предварительной повестки дня

Сбор данных через Интернет

Информационная вебсистема итальянской переписи населения

Записка ИСТАТ, Италия

Резюме

В ходе пятнадцатой переписи населения Италии впервые были применены некоторые инновационные решения, поддерживающую разнообразную деятельность в рамках процесса сбора данных и включающие в себя использование и анализ местного регистра населения (*Anagrafe*), привлечение подразделений местных муниципалитетов к полному участию в процессе переписи, а также стратегию многоканального сбора данных. В настоящем документе описаны основные характеристики вебсистемы.

I. Введение

1. Настоящий документ подготовлен Маурой Джакуммо, Леонардо Тинини, Мариной Вентури и Антонино Вирджилито из Национального института статистики Италии (ИСТАТ).
2. В ходе пятнадцатой переписи населения Италии впервые были применены некоторые инновационные решения, поддерживающие разнообразную деятельность в рамках процесса сбора данных. Первое решение касалось стратегии переписи, основанной на использовании и анализе местного регистра населения (*Anagrafe*), который использовался в качестве информационной базы пе-

реписи для сравнения и согласования данных о гражданах, собранных в ходе переписи, с данными местных регистров населения. Второе решение касалось привлечения ко всему процессу переписи подразделений местных муниципалитетов, которые участвовали в каждом этапе переписи, начиная со сбора и кончая мониторингом данных и регистрацией основных переменных для подготовки публикации предварительных данных, а в некоторых случаях – даже по всему вопроснику. Третье и, возможно, самое важное инновационное решение касалось методов многоканального сбора данных. Традиционный метод сбора данных счетчиком путем их записи на бумаге в ходе непосредственного опроса населения сочетался с электронным вопросником, который может использоваться как через Интернет для самостоятельного заполнения его гражданами, так и сетевым персоналом для ввода данных, первоначально записанных на бумаге. В настоящем документе описаны основные характеристики вебсистемы, которая состоит из трех основных программ, интегрированных в единую платформу: Системы управления переписью (СУО), онлайн-вопросник (ОВ) и онлайн-документации для операторов (ОДО).

3. СУО доступна как для сотрудников переписного управления, так и для персонала ИСТАТ. Она предназначена для использования на различных стадиях сбора данных. В частности, она позволяет сотрудникам переписного управления руководить процессом переписи путем распределения домохозяйств между счетчиками, контроля сбора данных и их просмотра. Поскольку сбор данных осуществляется по нескольким каналам, СУО позволяет отслеживать вопросники, собранные различными способами (Интернет, счетчики, почтовые/муниципальные центры сбора). СУО используется также статистиками для сбора основных переменных с целью включения их в предварительные результаты переписи. И наконец, еще один базовый модуль СУО позволяет управлениям местной администрации сравнивать и согласовывать данные о гражданах, собранные в ходе переписи, с данными из местных регистров населения.

4. Система онлайн-вопросника позволяет итальянским гражданам заполнять свои вопросники через Интернет. Интерфейс помогает пользователям следовать правилам заполнения и проверять ошибки перед окончательным представлением вопросника. После заполнения вопросник сразу же поступает в СУО. Онлайн-вопросник может использоваться также сотрудниками переписного управления для ввода данных. Выбор единой вебплатформы, включающей как онлайн-вопросник, так и систему управления, обладает такими преимуществами, как оперативный мониторинг процесса, улучшение качества данных, уменьшение объема работы на местах, сокращение лага между сбором данных и обработкой данных. Данная система обеспечила эффективное взаимодействие между ИСТАТ и Сетью переписи населения. В этом контексте в операциях по переписи населения были задействованы около 100 000 операторов. Более 30% граждан заполнили свои вопросники через систему ОВ, что свидетельствует о востребованности этого канала.

II. Система управления переписью населения (СУО)

5. Для поддержки функций переписной сети, касающихся управления различными стадиями переписи населения и жилищного фонда, была внедрена система, работающая на основе информационных технологий. Система управления и мониторинга позволяет муниципальным отделам контроля и учета отслеживать вопросники, полученные по разным каналам сбора данных, и руко-

водить работой счетчиков на местах, включая сбор недостающих вопросников и решение проблемы неполного охвата и избыточного охвата.

6. Описанные во введении инновации потребовали дополнить обычную систему мониторинга некоторыми функциями, такими как синхронизация и обмен данными с внешними системами; синхронизация с вебсистемой сбора данных; повышение сложности функций мониторинга в связи с диверсификацией вопросников и возможностью их возврата муниципалитетам по различным каналам; использование информации для руководства ежедневной и систематической работой счетчиков. Для этого было создано специальное приложение, основанное на использовании вебтехнологий: *CYO (Sistema di Gestione della Rilevazione* – система управления обследованием). Эта программа обеспечивает максимальную безопасность данных при их передаче и хранении в соответствии со стандартными правилами Национального института статистики.

7. Систему управления можно рассматривать как систему организации последовательных бизнес-операций, в которой каждый оператор может работать автономно, следуя четко установленной процедуре. Кроме того, такая схема позволила управлять технологическими процессами (удалять вопросники, менять статус и т.д.), что дало ИСТАТ возможность предупреждать проблемы, которые можно было решать лишь путем ручного вмешательства в отдельные вопросники.

8. Такой рабочий процесс обеспечил выгоды с точки зрения своевременности, качества данных и расходов. Система оказалась способна реализовать переписное обследование более 25 000 000 домохозяйств, или более 60 000 000 отдельных лиц. Кроме того, системой могли одновременно пользоваться более 100 000 операторов (переписной сети), которые работали ежедневно по несколько часов в день. Система имеет более 70 функций, сгруппированных по типам и организованных в 8 макрокатегорий.

9. **Изменения в местном регистре населения.** Эта категория касается небольших муниципалитетов с населением (менее 20 000 человек), которые могут вносить в систему изменения между списками местного регистра населения, направленными в ИСТАТ 30 декабря 2010 года, и списками по состоянию на 9 октября 2011 года (дата переписи). Специальные формы позволяют вносить информацию об изменениях; данные о новых домохозяйствах, о семьях, переехавших в муниципальный округ или покинувших его. Другие муниципалитеты направляли свои местные регистры населения через специальную вебпрограмму, и их данные загружались в СЮО.

10. **Операторы.** Эта группа функций позволяет муниципалитетам автономно управлять вспомогательным и полевым персоналом. Пользователи, уполномоченные использовать эти функции, могут вводить, просматривать и изменять персональную информацию о координаторах и счетчиках. Каждый сетевой сотрудник, который использовал систему, был зарегистрирован своим руководителем через специальные регистрационные формы. Эта операция была необходима для создания профиля пользователя и пароля, позволяющего войти в систему. Для каждого пользователя был также создан профиль, позволявший ему видеть только те функции, которые соответствуют его сфере компетенции. Из соображений безопасности пароль высылался на указанный при регистрации адрес электронной почты, и его необходимо было поменять при первом входе в систему. Новый паспорт сохраняется в системе в зашифрованном виде. Уполномоченные операторы могут также распределять счетчиков между координаторами, а переписные участки (и вопросники) между счетчиками.

11. **Краткие отчеты.** Включают набор отчетов, позволяющий постоянно следить за ходом обследования и работой счетчиков. Эти отчеты позволяют оценивать работу полевых операторов "практически в режиме реального времени". Этот мониторинг призван был интегрировать различные типы данных, которые связаны с многоканальным характером анкетирования и различными версиями вопросника (в длинной и краткой форме). Кроме того, в зависимости от уровня ответственности информация предоставлялась с разной степенью детализации благодаря поисково-аналитическому механизму, с тем чтобы разбивать элемент данных на максимальный уровень детализации, представленной муниципалитетом или счетчиком.

12. **Строения.** Эта категория относится к переписи жилищного фонда и включает регистрационную форму и отчеты о мониторинге. Для того чтобы помочь счетчикам передвигаться по муниципалитетам, система позволяет скачивать документы в формате PDF со всеми адресами муниципалитетов, поделенных на переписные участки.

13. **Вопросники.** Эта категория охватывает все функции, непосредственно связанные с обследованием; система предоставляла форму для регистрации вопросников, которые возвращали муниципалитеты, с целью проведения тщательного мониторинга такой информации. Уполномоченные муниципалитеты могли вводить данные, используя ту же вебпрограмму, которую могли использовать домохозяйства. Для сбора основных переменных, необходимых для распространения данных, система предоставляла специальную функцию ввода данных, доступную для муниципалитетов, которые не имели права вводить данные. Операторы могли распечатывать любой онлайн-вопросник.

14. **Переписной участок.** Эта категория охватывает все функции, которые управляют работой счетчика на местах. Основным элементом этой категории является программа переписи на уровне участка, которая отслеживает изменения в статусе каждого вопросника. Изначально она содержит данные из регистров населения и данные с неполным охватом, выявленные на предыдущем этапе, а также список всех адресов по каждому переписному участку. В эту программу постоянно вводятся новые данные из системы сбора данных, а также сведения от счетчиков и операторов отделов учета и контроля.

15. **Перепись и регистр населения.** Представляет собой информационную панель, позволяющую уполномоченным операторам отслеживать и устранять различия между данными регистра населения и результатами переписи. Каждый вопросник подвергается обработке, и операторы отмечают, совпадают ли лица, указанные в вопроснике, с теми, которые значатся в государственном регистре, или, в случае необходимости, добавляют новых лиц. По окончании этой работы система выдает четыре разных отчета: а) граждане, живущие в одном жилище, б) граждане, живущие в другом жилище, с) граждане, включенные в вопросник, но не включенные в государственный регистр, d) граждане, включенные в государственный регистр, но не включенные в вопросник.

16. **Утилиты.** Данная категория включает набор функций поддержки сетевого режима, охватывающих весь процесс обследования.

III. Онлайн-вопросник (ОВ)

17. Онлайн-вопросник (ОВ) итальянской переписи населения представляет собой предлагаемое гражданам вебприложение, позволяющее им заполнять свои переписные листы точно таким же образом, как они заполняют пе-

чатные формуляры. На первой странице печатного вопросника указывался код авторизации, который вместе с индивидуальным номером налогоплательщика может использоваться респондентом в качестве пароля для входа в данную веб-систему. Эта программа воспроизводит все три вида печатных вопросников, соответственно два вопросника для семей и домохозяйств (в длинном и кратком варианте) и вопросник для сожителей/сожителей.

18. К концу сбора данных через систему ОВ было направлено почти 8 400 000 заполненных вопросников, что составляет 33% от их общего прогнозируемого количества. В среднем в первые два месяца процесса переписи в минуту приходило по 100 вопросников, а в самые активные периоды направлялось до 300.

19. Система ОВ была успешно интегрирована в СУО: после заполнения респондентом вопросника в программе ОВ в центральной базе данных сразу же обновлялся статус этого вопросника, который становился видимым для операторов через СУО. Операторы могли также использовать ОВ для ввода данных.

20. Поскольку система ОВ была потенциально доступна всему населению Италии, крайне важно было тщательно ее спроектировать таким образом, чтобы она могла функционально подходить для огромного числа пользователей и при этом была очень надежной (т.е. пользователи не должны сталкиваться с ошибками в приложении) и обеспечивала защиту данных, несмотря на все возможные угрозы безопасности. При внутреннем проектировании программы использовались современные технические решения для: i) уменьшения нагрузки на базу данных даже при одновременном использовании программы многочисленными пользователями, и ii) обеспечения проверки подлинности данных на многочисленных этапах работы программы, с тем чтобы гарантировать целостность данных, направляемых пользователем.

21. Кроме того, в этой программе активно используются метаданные, которые позволяют обойтись без разработки резервного программного кода для трех вопросников. Все, что появляется на веб-интерфейсе пользователя (включая предупреждения и сообщения об ошибках), хранится в специальных таблицах метаданных и ресурсных файлах. В частности, текст каждого отдельного вопроса переписного листа хранится в таблице базы (мета)данных с тремя разными локализованными версиями (итальянской, немецкой и словенской). Модели единственного возможного ответа также хранятся в таблице базы (мета)данных с тремя различными локализованными версиями. В любой момент можно было вносить изменения и поправки в текст вопросников (что и делалось), даже после окончательного ввода программы в эксплуатацию и предоставления гражданам доступа к ней, без изменения исходного кода программы.

22. Тексты не являются единственным элементом программы ОВ, хранящимся в виде метаданных. Еще одна важная особенность ОВ заключается в том, что некоторая часть "поведения" программы также хранилась в метаданных. Это тесно связано с концепцией "графа вопросника", который является основополагающей частью этой технологии, используемой для формального моделирования структуры вопросника и правильного набора и правильной последовательности вопросов, на которые должен ответить респондент.

23. Граф вопросника представляет собой ациклический оргграф (АО), в котором: i) вершины находятся в соответствии 1–1 с каждым вопросом переписного листа; ii) (ориентированное) ребро от вершины (вопроса) N_i к вершине (вопросу) N_j соответствует тому, что пользователь должен ответить на вопрос N_j после того, как он дал ответ для вершины N_i . Однако в общем от одной и той же вер-

шины N_i может исходить более одного ребра (например, "Если Вы ответили "Да", перейдите к вопросу X.Y, в противном случае продолжайте отвечать на вопросы"). Следовательно, у вершин могут быть метки, т.е. условия, зависящие от ответов на вопросы. Условия на ребрах, выходящих из любой конкретной вершины N_i , должны быть взаимоисключающими (и одно из них обязательно верным) и определять, какой из вопросов действительно является "следующим" после N_i .

24. Граф вопросника, т.е. данные о вершинах, ребрах (и связанных с ними условиях), хранится в базе (мета)данных ОВ и используется программой (как с клиентской, так и с серверной части) для визуального включения и отключения вопросов на вебстранице и для проверки вводимых пользователем данных перед сохранением ответов пользователя в таблицах микроданных. Поскольку граф является инвариантным, его структура "кэшируется" в основной памяти программой ОВ во избежание ненужных и неэффективных повторяющихся запросов к базе данных.

25. На основании графа вопросника и уже представленных пользователем ответов программа автоматически определяет, какие вопросы необходимо включить и отключить на вебинтерфейсе пользователя. Благодаря тому, что граф вопросника является на деле АО, алгоритм обновления включения/отключения вопросов особенно эффективен: в частности, может быть показано, что каждое ребро должно быть рассмотрено лишь один раз для обновления свойства включения вопросов по всему переписному листу. Если пояснить в двух словах, то в основе алгоритма лежит следующая идея: с учетом уже представленных пользователем ответов вопрос должен быть включен, если пользователь обязательно дойдет до него, независимо от ответов, которые он еще не дал. И наоборот, он должен быть отключен, если пользователь точно не дойдет до него или же существует последовательность ответов на оставшиеся еще без ответа вопросы, которая не позволит пользователю дойти до этого вопроса.

26. В системе ОВ используется также передовой механизм автоматического кодирования текстовых ответов на основе словаря, сравнения схожих строк и автоматического ранжирования. Этот механизм использовался для вопроса об уровне образования. Соответствующий классификационный словарь состоит из более чем 6 000 различных записей, и требовалось, чтобы система ОВ представляла уже закодированное название (иными словами, система ОВ не могла хранить текст на естественном языке, который впоследствии необходимо было бы закодировать вручную или полуавтоматически). Выбор ответа из выпадающего списка естественно невозможно было использовать из-за большого количества возможных вариантов.

27. Следовательно, был выбран механизм, очень похожий на взаимодействие, возникающее в поисковых системах: пользователь вводит набор слов, описывающих полное название его образовательной квалификации; система отвечает, предлагая перечень названий, взятых из существующего словаря и ранжированных по сходству с тем, что указал пользователь. И наконец, пользователь может выбрать один из предложенных вариантов или начать новый поиск, если предлагаемый перечень его не устраивает.

28. Для повышения эффективности поисковой системы и во избежание перегрузки системных серверов осуществляется предварительная обработка словаря. Эта предварительная обработка представляет собой достаточно сложную с вычислительной точки зрения специальную процедуру, и осуществляемую по этой причине в режиме оффлайн. Эта процедура состоит из следующих основных этапов: i) нормализации знаков в словарных записях (например, буквы с

диакритическими знаками заменяются соответствующими буквами без них); ii) удаления "игнорируемых слов" (например, артиклей, союзов и других "бесполезных" слов) из словарных записей; iii) извлечения отдельных слов из нормализованных и упрощенных вариантов словарных записей, получившихся в результате предыдущих шагов. Эти слова (называемые в дальнейшем "слова-дескрипторы") хранятся в базе данных и связаны с исходными вариантами словарных записей для последующего использования онлайн-поисковой системой.

29. Онлайн-поисковая система получает в качестве данных (вводимых пользователем) исходную поисковую строку, состоящую из нескольких слов, разделенных пробелами. Сначала поисковая строка нормализуется, упрощается и делится на набор отдельных поисковых слов, аналогично тому, что было сделано в отношении словарных записей. Затем осуществляются следующие шаги: i) каждое поисковое слово сравнивается со словами-дескрипторами из базы данных: если соответствие выше данного порогового значения, то слово-дескриптор добавляется в перечень Lt слов, превышающих пороговое значение; ii) все словарные записи, связанные со словами из перечня Lt, извлекаются из базы данных; iii) записи, извлеченные на предыдущем этапе, сортируются в соответствии с несколькими критериями, в частности, такими как: точное или близкое совпадение, частотность каждого слова в словаре, количество совпадающих слов, общее количество слов в данной записи.

IV. Сетевой портал (ОДО)

30. Сетевой портал представляет собой главным образом информационный вебсайт, организованный в виде двух горизонтальных меню навигации, которые предоставляют доступ к различным информационным областям. Существует 7 информационных областей, которые различными путями оказывают помощь оператору.

31. Область **документов сбора данных** содержит документы, которые также являются рабочими инструментами для проведения переписи. Кроме вопросников в формате PDF, касающихся семьи и жилищного фонда, в этой области также содержатся пользовательские инструкции и документация, способные оказать помощь в составлении некоторых разделов переписного листа (например, вспомогательные средства и коды).

32. Область **инструментов** содержит ПО, помогающее гражданам выбрать правильную экономическую и трудовую деятельность, а область **документов** содержит официальные документы по связанным темам, такие как основные законы о населении и защите персональных данных.

33. Другие области представляют собой раздел **"вопросы и ответы"** (часто задаваемые вопросы), **гlossарий**, **обучающие видеопособия и программы**, которые содержат учебные материалы для переписной сети, в том числе интерактивную версию вопросника, пользовательские инструкции, разделенные на главы, и различные слайды по содержанию аудиторных курсов обучения.