

Распространение: Общее
23 апреля 2012 г.

Оригинал: английский
(неофициальный перевод)

Европейская экономическая комиссия

Конференция статистиков Европы

Группа экспертов по переписи населения и жилого фонда

Четырнадцатое заседание

Женева, 24-25 мая 2012 г.

Пункт 5 предварительной повестки дня

Сбор данных через Интернет

Сбор данных через Интернет в рамках переписи населения и жилищного фонда 2011 года в Чешской Республике

Записка Статистического органа Чешской Республики¹

Резюме

Перепись населения на территории современной Чешской Республики проводится уже несколько столетий. Но впервые в рамках переписи населения и жилищного фонда 2011 г. была предоставлена возможность использования электронных переписных листов и Интернета. Все жители и собственники домов и квартир могли принять решение использовать бумажные или электронные переписные листы. При первом посещении интервьюера каждое домохозяйство получило комплект бумажных переписных листов. На каждом переписном листе был указан специальный PIN-код для того, чтобы открыть в Интернете электронный переписной лист.

Вся система электронных переписных листов была разработана и управлялась внешней информационной компанией. Поставщик получил основные требования к системе в части пропускной способности и безопасности. Обеспеченная фактическая пропускная способность электронной системы ни разу не использовалась более чем на одну треть одновременно, и не произошло ни одного инцидента безопасности.

Все пользователи могли использовать бумажные и электронные переписные листы в одно время в течение четырех недель. Это легло в основу просветительской

¹ Автор- Stanislav Drápal.

кампании. Электронные переписные листы были применены 27,5% жителями, а доля электронных переписных листов в составе всех заполненных переписных листов составила 25,5%.

I. Электронные формы

1. При разработке сложных решений для электронных систем сбора данных критически важными элементами являются тип электронных форм и технология, применяемая в рамках системы электронных форм. Как и традиционные бумажные формы, электронные формы применяются для сбора различных видов данных. Важно, чтобы электронные формы внешне были очень похожи на бумажные формы. Для того чтобы правильно заполнить электронные формы, предлагаются разнообразные интерактивные элементы (текстовые поля, кнопки с независимой фиксацией, клавиши с изменяемой функцией и т.д.). Их внешний вид и функциональное назначение очень близки к тому, что пользователи уже знают из своей повседневной деятельности. Кроме того, включаются преимущества электронной обработки (например, электронные формы могут быть автоматически заполнены некоторыми данными, полученными из других источников и т.д.).

2. Другими важными элементами решений для электронных форм являются технологии, необходимые для распространения форм, а также их сбора. К таким технологиям относятся, например, представление электронных форм на заданном веб-сайте, проверка данных, введенных пользователями, взаимодействие с пользователем, передача данных в базы данных или другие хранилища данных, подтверждение, отправляемое отправителю, и многие другие.

II. Электронные формы, их виды и технологические основы

3. Было решено, что наилучшим пересечением между функционалом и открытостью является технология Adobe, чьи электронные формы традиционно были связаны с форматом PDF и приложением Adobe Acrobat. Используя приложение можно создавать формы, чьи элементы управления и интерактивность сравнимы или даже лучше предлагаемых формами HTML (например, вывод на печать, сохранение или автономность, которые делают электронные формы равноценными бумажным формам). Кроме того, заслуживала внимания возможность применения форм в формате PDF в сочетании с технологиями для электронной подписи и безопасности. Более того, возможности PDF по представлению информации лучше тех, которые обеспечивает HTML (а также другие технологии, связанные с формами), что делает платформу PDF еще более интерактивной. И, наконец, приложение, которое необходимо для того, чтобы видеть и заполнять формы PDF, мультиплатформа Adobe Reader, имеется в свободном доступе.

III. Решения для электронных форм

4. Во время переписи населения и жилищного фонда 2011 г. для распространения и сбора электронных форм использовалось решение, основанное на технологии форм Adobe PDF. Причиной выбора этого решения при сравнении с онлайн-формами HTML являлось комплексное выполнение следующих требований:

- Минимизированный поток информации – во время исследования, проведенного до переписи, была выявлена значительная заинтересованность респондентов² в онлайн-представлении форм.
- Независимость форм – каждая форма должна содержать все функции и информацию с тем, чтобы респонденты могли правильно заполнить свои ответы, без необходимости какого-либо взаимодействия с центральными системами.
- Идентичный вид – Имело место требование о том, что электронная форма и бумажный переписной лист должны быть одинаковыми с тем, чтобы общественность могла легко понять инструкции и заполнить данные. Респонденты получают одни и те же инструкции перед тем, как принять решение о том, какую форму переписного листа использовать.
- Доказательное происхождение – электронные формы должны содержать в себе механизм для доказательства своего происхождения, другими словами, для демонстрации того, что этот переписной лист предоставлен Статистическим органом Чешской Республики.
- Технологическая последовательность – технология, применяемая населением для заполнения переписных листов, должна быть широко известной, бесплатной и обеспечиваться стандартными платформами для того, чтобы население могло последовательно ее использовать.
- Политика защиты конфиденциальности частной информации – технология должна поддерживать защиту личных данных, предоставляемых респондентами, и должно четко указываться, что данные не сохраняются в компьютере без ведома респондента. Данный подход открывает возможность использования компьютеров, имеющихся для открытого доступа в учебных комнатах, библиотеках и муниципальных органах для граждан, которые не могут использовать собственные компьютеры.
- Подтверждение подачи – решение должно включать функциональную возможность для подтверждения того, что респондент заполнил переписной лист, которое может быть сохранено в электронной форме или распечатано и не может быть подделано.

5. Таким образом, решение, основывающееся на технологии Adobe PDF, было разделено на следующие сегменты, которые готовились отдельно, а в конце частичные решения были интегрированы в целях формирования комплексного решения.

- Подготовка переписных листов – разработка документов PDF, которые представляли собой конкретные виды переписных листов, включая необходимые функциональные возможности и опции и возможную итоговую проверку респондентом.
- Предварительное наполнение переписных листов – подготовка механизма, который будет предварительно наполнять переписные листы респондентов в соответствии с бумажными переписными листами.
- Представление информации – представление информации, связанной со скачиванием, заполнением и подачей переписного листа.

² Примерно 10 % респондентов

- Распространение переписных листов – электронный переписной лист имеется только по респондентам, данные которых заполнены.
- Сбор переписных листов – сбор данных, представленных путем заполнения формы, и подтверждение подачи данных.
- Техническая инфраструктура – развитие необходимой технической инфраструктуры в целях обеспечения распространения и сбора в соответствии с требованиями безопасности.

IV. Подготовка переписных листов

6. Электронные переписные листы были подготовлены в виде отдельных документов PDF для каждого типа формы, т.е. лист переписи населения, лист переписи зданий и лист переписи жилищного фонда. Переписные листы были разработаны таким образом, чтобы они выглядели идентично с бумажными версиями. Помимо унификации информации, предоставляемой респондентам в обеих формах (электронной или бумажной), такая схожесть охватывала ситуацию, когда пользователь распечатывал электронный переписной лист, заполнял его и подавал его как бумажный переписной лист: идентичный вид позволял оцифровывать данные.

7. Переписные листы включали программный код для оказания контекстуальной помощи при заполнении каждого поля. Кроме того, код содержал проверку данных на уровне конкретных полей (например, вы не можете ввести буквы в поле, где необходимо ввести цифры), затем был логический контроль по обусловленным полям (например, пол и количество детей – мужчина не может родить ребенка) и в случае с некоторыми конкретными полями предлагался выбор значений для заполнения в форме. Перед отправкой данных в Статистический орган проводилась автоматическая проверка на полноту данных, заполненных в форме. Вышеуказанные проверки значительно повысили качество данных, собираемых посредством электронных переписных листов.

8. Формы подписывались электронным сертификатом, выдаваемым для Статистического органа VeriSign, очень известным сертификационным органом. Общественность была в достаточной степени детально проинформирована о том, каким образом автор документа может быть проверен в приложении Adobe Reader. Таким образом обеспечивалась бескомпромиссность и надежность загружаемого переписного листа, устраняя опасность атак.

9. В отношении требований к функциональным возможностям, использовались документы PDF, разработанные для относительно нового приложения Adobe Reader. Решение по использованию технологий Adobe оказалось верным, поскольку у респондентов не возникло каких-либо проблем при обновлении или установке Adobe Reader, что доказало преимущества универсального инструмента по сравнению с (ранее рассматривавшимися) решениями, основывающимися на HTML или X-формам.

V. Предварительно заполняемые данные в переписных листах

10. С учетом того, что бумажные переписные листы предоставлялись респондентам включая данные, уже известные из административных источников, для того, чтобы упростить процесс заполнения и повысить качество собираемых данных,

такая концепция была применена и в отношении электронных переписных листов. Кроме того, каждый электронный переписной лист был однозначно предназначен для определенного респондента.

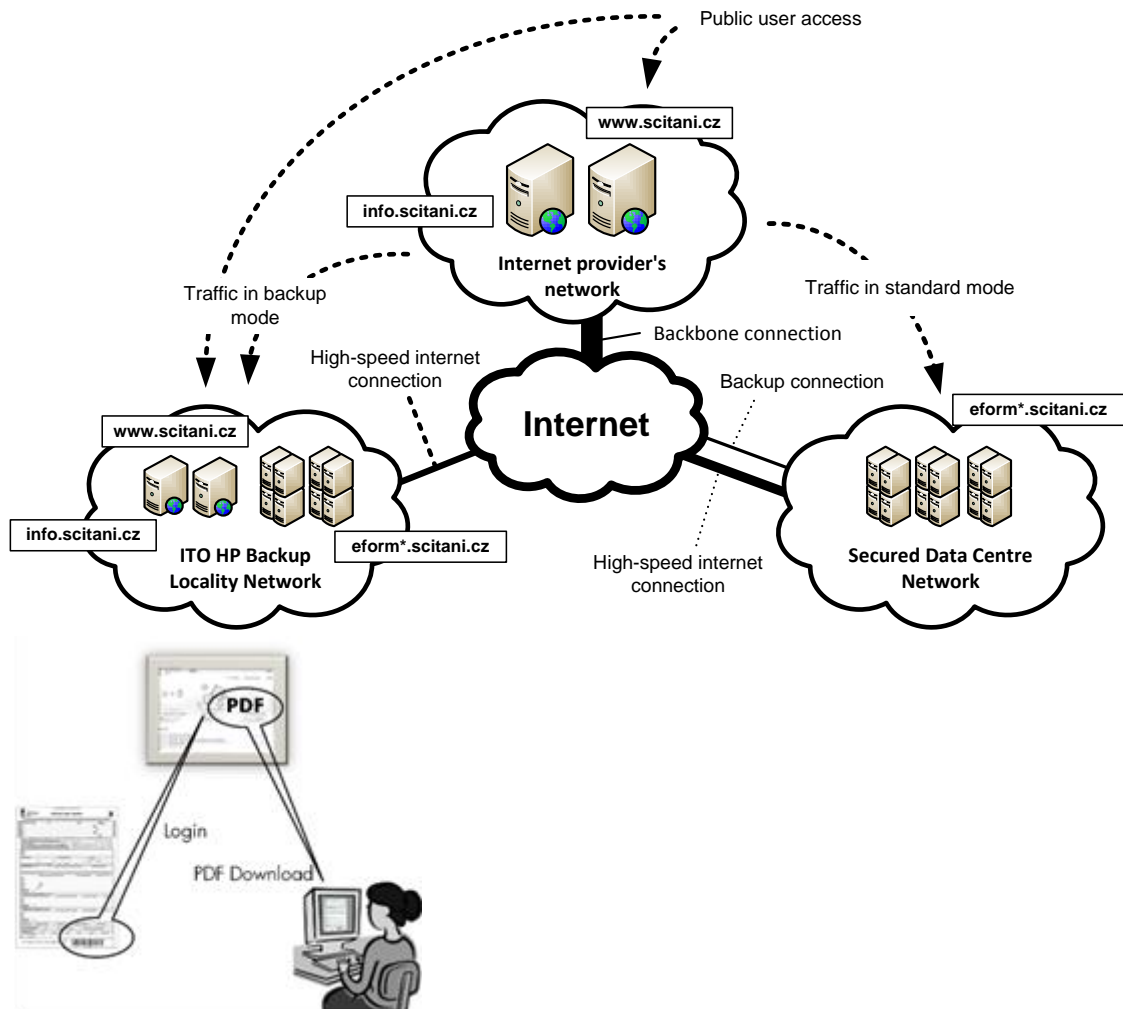
11. В целях экономии времени электронные переписные листы были сгенерированы на этапе распространения, поскольку предварительное наполнение началось до переписи. Около 20 млн.³ переписных листов было предварительно заполнено и сохранено в базе данных емкостью 4 Тб для последующего распространения. Работа по предварительному наполнению данных заняла несколько дней.

VI. Представление информации

12. На этапе подготовки к переписи Статистическим органом была запущена специальная веб-презентация для общественности, содержащая информацию о распространении, заполнении и сборе заполненных переписных листов.

13. Ожидалось, что общественность будет заинтересована в онлайн-переписи, поэтому презентация была разделена на две части – внешний (frontend) сервер, называющийся “signpost” (www.scitani.cz) с общей информацией, и второй сервер, содержащий детальную презентацию (info.scitani.cz).

³ Не было известно кто из респондентов воспользуется электронным переписным листом, поэтому данные были заранее загружены во все формы.



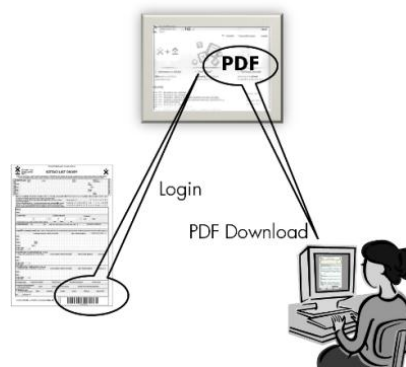
14. Две части представлялись отдельно с использованием внешней инфраструктуры с несколькими местоположениями в магистральной Интернет-сети, администрируемых разными Интернет-провайдерами (см. область на рисунке, обозначенную как «сеть Интернет-провайдера»). Вследствие необходимости обеспечения защиты личных данных для распространения и сбора переписных листов использовалась инфраструктура Статистического органа.

15. Распространение и сбор осуществлялись из двух географически разных мест. Вследствие вышеуказанного механизма нагрузка на технические устройства, используемые Статистическим органом, была снижена (область на рисунке, обозначенная как «защищенная сеть центра данных») и была разработана распределенная инфраструктура (область, обозначенная как «сеть резервного места ИТО НР») для предоставления информации, с надежной базой для защиты системы от отказов и атак хакеров, нацеленных на презентации.

VII. Распространение переписных листов

16. В части решения для распространения электронных переписных листов ключевой вопрос заключался в том, чтобы обеспечить, что человек будет скачивать форму с заранее загруженными данными, относящимися только к нему, его квартире или дому.

17. Было решено, что соответствующий механизм будет основываться на том, что каждый человек получит бумажный переписной лист. Форма содержала идентификационный номер и защитный код. Оба кода выбирались произвольно из соответствующего набора кодов. Набор был достаточно большим, поэтому применяемые коды были действительно произвольными и разбросанными. Так можно было избежать предсказания неизвестных кодов на основании уже известных кодов.

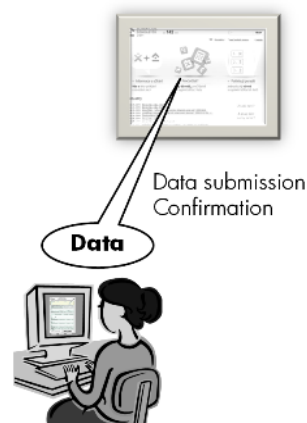


18. Когда человек хотел подать переписной лист в онлайн, он использовал коды из бумажного переписного листа, вводил их в Интернет-приложении, загружал электронные эквиваленты бумажных переписных листов на свой компьютер и заполнял данные.

VIII. Сбор переписных листов

19. С учетом того, что размер данных, предоставляемых респондентом в форме, был в сотни раз меньше емкости форм Adobe PDF, было решено, что Статистический орган будет получить не заполненную форму, а только данные, предоставляемые респондентами.

20. Для этого в приложение Adobe Reader была включена функциональная возможность, позволяющая отправлять данные как документ XML посредством защищенного (HTTPS) протокола. Таким образом, данные получались в виде коротких XML сообщений для дальнейшей обработки.



21. После отправки и успешного получения данных приложение просило загрузить другой документ PDF, содержащий электронное подтверждение подачи данных. Это было реализовано путем отправки произвольного кода подтверждения, который был привязан к конкретному переписному листу. Респондент мог сохранить подтверждение на компьютере или распечатать его.

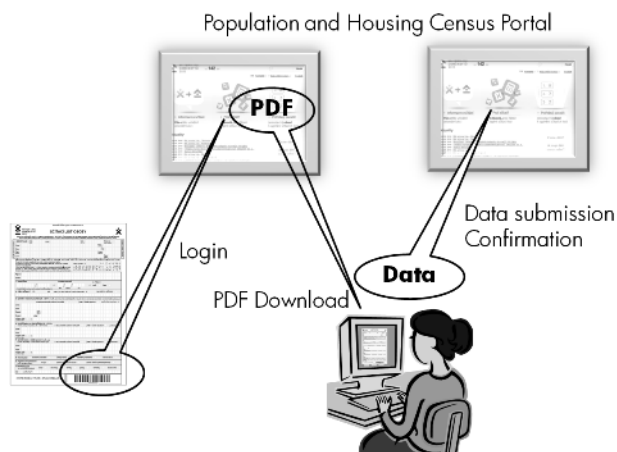
IX. Техническая инфраструктура

22. Техническая инфраструктура для распространения и сбора электронных переписных листов была размещена в трех местах:

- i. Основная техническая инфраструктура для приложений, предназначенных для загрузки электронных переписных листов и сбора данных из переписных листов, функционировала в защищенном центре данных, арендованном Статистическим органом для проведения переписи населения и жилищного фонда.
 - ii. Резервная техническая инфраструктура для приложений загрузки электронных переписных листов и сбора данных респондентов обслуживалась генеральным поставщиком информационных услуг посредством защищенного сервиса распределенной сети с ретрансляцией/коммутацией пакетов.
 - iii. Инфраструктура для презентаций для общественности была реализована посредством аренды у Интернет-операторов технологий в магистральной сети.
23. В части безопасности ключевым элементом являлась основная техническая инфраструктура. Она включала отдельные защищенные зоны в соответствии со стандартной архитектурой так называемых «демилитаризованных зон», которые образуют безопасную буферную область между частью архитектуры, доступной для общественности, и закрытой частью архитектуры.
24. Базы данных и демилитаризованные зоны были соединены только посредством проходов, инициируемых с уровня базы данных. Таким образом, собранные данные извлекались в «псевдо-онлайновом режиме» из демилитаризованной зоны во внутренние базы данных посредством передачи данных.
25. Внешний интерфейс сервиса по распространению и сбору переписных листов включал комплекс из пяти серверов. Один из серверов предназначался для работы с запросами, поступающими от IP-адресов, ассоциирующихся с высокой вероятностью атак DDoS (распределенная атака типа «отказ в обслуживании»), в основном из Азии. В случае успешной атаки такого типа отказал бы только один сервер, а все остальные серверы продолжали бы сервис по распространению переписных листов и сбору данных. Поскольку перепись населения и жилищного фонда проходила на территории Чешской Республики, доступ с вышеуказанных IP-адресов должен был быть практически нулевым. Более высокий трафик с этих IP-адресов означал бы попытку атаки.
26. Другим уровнем безопасности в части несанкционированного доступа была деятельность, связанная с внешним межсетевым экраном, поддерживаемым IDS (система обнаружения атак), основанной на технологической платформе HP Tipping-Point. Цель заключалась в обеспечении проактивного обнаружения атак на оказываемый сервис.
27. Резервная техническая инфраструктура следовала по своей архитектуре основной инфраструктуре, за исключением того, что использовались виртуальные технические элементы (серверы и сетевые устройства). В целях снижения рисков, связанных с безопасностью, резервная инфраструктура держалась в режиме «горячего резервирования» для того, чтобы суметь перенаправить процессы на резервную точку размещения с минимальным воздействием в случае отказа или перегрузки основной инфраструктуры.

X. Краткое повторение с точки зрения пользователя

28. С точки зрения пользователя ключевыми элементами комплексного решения были безопасность, прозрачность и удобство, как с точки зрения комплексного распространения переписных листов и сбора данных, так и с точки зрения самого переписного листа.



29. В рамках указанных выше требований разработанная система была действительно очень простой. Это означает, что она была надежной и удобной в использовании и, с точки зрения пользователя, она предусматривала лишь несколько действий, как показано на рисунке выше.

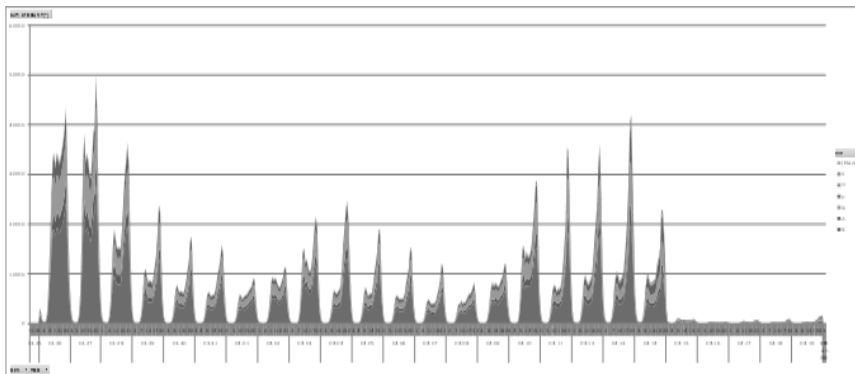
- i. Для каждого распечатанного бумажного переписного листа имелась электронная форма в формате PDF. Таким образом, на этапе распространения граждане лишь вводили номер и код безопасности со своего бумажного переписного листа, которому в базе данных соответствовал эквивалент в формате PDF, и конкретным гражданам предоставлялся конкретный переписной лист посредством защищенного соединения (HTTPS).
- ii. Затем человек заполнял необходимые данные в предварительно частично заполненном переписном листе в формате PDF, который выглядел практически так же, как и бумажная форма. Форма была обеспечена функциональными возможностями для проведения логической проверки значений, вводимых в выборочные поля, а значения в других полях, которые необходимо было заполнить, предлагались в виде «выпадающего» меню.
- iii. Проверенное содержание (не полный переписной лист), т.е. данные минимального размера по сравнению с размером переписного листа, отправлялось через защищенный канал в базу данных Статистического органа для дальнейшей обработки.
- iv. Получение данных из представленного переписного листа подтверждалось новым документом PDF, содержащим уникальное подтверждение подачи данных.

XI. Электронные формы - результаты и оценка

30. Во время переписи населения и жилищного фонда 2011 г. впервые граждане Чешской Республики могли заполнить электронные формы через сеть Интернет. В

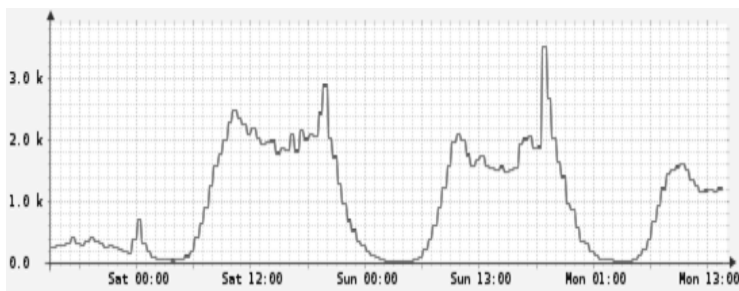
ходе переписи населения и жилищного фонда 2011 г. Статистический орган получил 4,33 млн. электронных переписных листов, которые включали 2,8 млн. личных страниц, 1,05 млн. страниц по домохозяйствам и 0,48 млн. страниц по жилищам. Доля электронных форм в составе всех поданных переписных листов составила 25,5%.

31. Наибольшее число электронных форм (509 тыс.) поступило в первый день сбора данных, в субботу 26 марта. Еще почти полмиллиона форм поступило в воскресенье 27 марта. В течение первой недели 44% всех представленных переписных листов было получено через Интернет. На следующем рисунке показаны колебания в частоте представления форм в течение всего периода сбора данных:



32. Количество подаваемых форм зависело от дней недели. Наибольшее количество электронных форм было получено в выходные дни (примерно 40% от общего количества полученных форм), в основном в воскресенья (968 тыс. электронных форм). И, наоборот, наименьшее количество электронных форм было получено в пятницы (200 тыс.).

33. Предпочитаемым временем в течение дня было вечернее время. Около 40% всех электронных форм было подано в период между 18 и 22 часами, в этот промежуток каждый час поступало более 300 тыс. Наибольшая активность наблюдалась в период между 20 и 22 часами, когда было получено 939 тыс. форм. Формы представлялись в любое время дня и ночи. На следующем рисунке показан типичный ход процесса подачи форм:



34. Например, в период с 1 до 2 ночи было подано почти 13 тыс. форм, в то время как в период с 4 до 5 часов утра было подано более 2 тыс. форм.

35. Практически половина личных страниц и страниц по домохозяйствам возвращалась в течение 10 минут после загрузки. В случае со страницами по жилью в течение 10 минут возвращалось более 90% переписных листов.

36. С точки зрения архитектуры успех в плане использования электронных переписных листов в переписи населения и жилищного фонда 2011 г. был достигнут благодаря нескольким ключевым факторам. В основном, такой успех был обусловлен применением технологии Adobe PDF, которая оказалась подходящей в части удовлетворения требований к сбору данных и широко принималась населением.

37. В части электронных форм и их возможного использования основные вопросы касались реальной заинтересованности в методе электронного наблюдения, его распределения по времени и влияния на техническую инфраструктуру. На этом этапе, безусловно, наиболее важной частью было планирование и работа со СМИ. Наибольшая нагрузка на системы наблюдалась в то время, когда в средствах массовой информации во время новостей в прайм-тайм сообщалось о правовой обязанности участия переписи и заполнении переписных листов. Однако благодаря контролируемой работе Статистического органа со СМИ это было предсказуемо.

38. Результаты мониторинга нагрузки на техническую инфраструктуру показывают, что при реализации задачи, подобной этой, эксплуатационные параметры вычислительных систем являются достаточными.

39. Система защита, примененная для распространения и сбора электронных форм, успешно нейтрализовала 3 760 различных попыток атак. Наиболее распространенным типом атаки был межсайтовый скриптинг (XSS), который является методом вторжения на веб-страницы посредством ошибок безопасности в скрипте. Кроме того, были обнаружены неуспешные попытки атаки DoS. В целом, прогнозируемые риски, связанные с безопасностью, не проявились, и общественность положительно отнеслась к электронным переписным листам. Люди осознавали, что эта услуга экономит их время и помогает облегчить выполнение их правовой обязанности.
