



**Экономический
и Социальный Совет**

1

Distr.
GENERAL

ECE/CES/GE.41/2007/7
26 March 2007

RUSSIAN
Original: ENGLISH/RUSSIAN

ЕВРОПЕЙСКАЯ ЭКОНОМИЧЕСКАЯ КОМИССИЯ

КОНФЕРЕНЦИЯ ЕВРОПЕЙСКИХ СТАТИСТИКОВ

Группа экспертов по переписям населения и жилищного фонда

Десятая сессия

Астана, 4-6 июня 2007

Пункт 3 (b) предварительной повестки дня

**ТЕХНОЛОГИИ ПРОВЕДЕНИЯ ПЕРЕПИСИ: ПОСЛЕДНИЕ ИЗМЕНЕНИЯ
ПРОВЕДЕНИЯ ПЕРЕПИСИ И ИХ ПОСЛЕДСТВИЯ ДЛЯ МЕТОДОЛОГИИ**

**Опыт использования сканеров при обработке
данных переписи населения 1999 года**

Представлен Агентством Республики Казахстан по статистике

Совещание проводится совместно с Евростатом

Резюме

Бюро Конференции европейских статистиков (КЕС) на своем совещании, состоявшемся в Вашингтоне, округ Колумбия, 19-20 октября 2006 года, одобрило обновленный круг ведения Руководящей группы по переписям населения и жилищного фонда и план деятельности КЕС в области переписей населения и жилищного фонда. Бюро КЕС также приняло решение о том, что Руководящая группа будет координировать работу по различным типам совещаний. Настоящий документ подготовлен по просьбе Руководящей группы по переписям населения и жилищного фонда для представления и обсуждения на Совместном совещании ЕЭК/Евростата по переписям населения и жилищного фонда, которое состоится в Астане (Казахстан) 4-6 июня 2007 года. Данный документ послужит основой для обсуждения на заседании, посвященном теме "Технологии проведения переписи: последние изменения в проведении переписи и их последствия для методологии".

I. Введение

1. Обработка данных Первой национальной переписи населения Республики Казахстан 1999 года является самым сложным, но в то же время успешным проектом, осуществленным в Информационно-вычислительном центре (ИВЦ) Агентства РК по статистике. При реализации проекта (в 1998-2001г.г.) использованы современные инструментальные средства для разработки прикладных программ и технология «клиент-сервер», подготовлены сильные специалисты по обработке данных. База данных, сформированная по итогам переписи не имеет аналогов в Казахстане и является одним из достояний Агентства по статистике. Предыдущие переписи населения проводились ЦСУ Казахской ССР и все данные направлялись в ЦСУ (Центральное статистическое управление) СССР, где формировались централизованные базы данных. Итоги этих переписей сохранены только в виде отпечатанных сборников. Обобщение итогов проведенной работы, анализ всего комплекса мер, реализованных при организации обработки материалов первой переписи населения и анализ ошибок очень важен именно сейчас, когда началась подготовка к проведению следующей переписи населения.

II. Организация кустовых центров обработки данных

2. По окончании проведения переписи (с 25 февраля по 4 марта 1999 года) все переписные материалы были проверены в инструкторских участках, далее в переписных отделах и сданы в районные отделы статистики. После дополнительной проверки переписные материалы были сданы в областные (городские) отделы статистики для кодирования, которые в свою очередь согласно графику по описи сдавали их в кустовой центр для последующей автоматизированной обработки. Всего было создано 5 кустовых центров:

- (a) Алматинский (ИВЦ Агентства РК по статистике) - для обработки данных Северо-Казахстанской, Алматинской областей и г.Алматы;
- (b) Актюбинский (г.Актобе) - для обработки данных Актюбинской, Атырауской, Мангистауской, Западно-Казахстанской и Кызылординской областей;
- (c) Восточно-Казахстанский (г.Усть-Каменогорск) - для обработки данных Восточно-Казахстанской и Павлодарской областей;
- (d) Карагандинский (г.Караганда) - для обработки данных Карагандинской, Акмолинской, Костанайской областей и г. Астана;

- (е) Южно-Казахстанский (г.Шымкент) - для обработки данных Южно-Казахстанской и Жамбылской областей.

3. Центр в ИВЦ выполнял также роль республиканского центра, т.е. все данные из кустовых центров передавались в ИВЦ для дальнейшей обработки. Все переписные бланки после осуществления оптического считывания возвращались из кустовых центров в управления статистики областей для дальнейшего хранения.

4. Материалы для обработки начали поступать в центры начиная с конца апреля месяца, ввод данных во всех центрах был завершен в конце августа 1999 года. Основные данные по объему обработанного материала приведены в Таблице 1. Можно отметить, что количество бланков введенных за один день через один сканер составляло от 15000 до 20000 штук.

Таблица 1. Основные показатели по объему обработанных данных

Центры обработки	Кол-во сканеров, <i>шт</i>	Кол-во переписных бланков, <i>шт</i>	Кол-во портфелей <i>шт</i>	Кол-во бланков обработанных одним сканером, <i>шт</i>	Продолжительность обработки, <i>дни</i>	Обработано бланков за один день, <i>шт</i>
Алматинский	3	4 743 300	15 800	1 581 100	102	46 503
Актюбинский	2	3 402 191	11 420	1 701 096	110	30 929
Восточно-Казахстанский	2	3 000 123	10 040	1 500 062	90	33 335
Карагандинск.	2	4 590 000	14 027	2 295 000	120	38 250
Южно-Казахстанский	2	3 710 657	12 847	1 855 329	90	41 230
Всего	11	19 446 271	64 134			

5. Для проведения работ по вводу и первоначальной корректировке данных были привлечены временные работники, общее количество которых составило 467 человек (см. **Таблица 2**). Уровень среднемесячной заработной платы работников центров обработки данных составлял 6,000 тенге.

Таблица 2. Количество работников, участвовавших в обработке данных

Центры обработки	Коррек- тировщики	Операторы по подготов- ке бланков для сканера	Специалисты технического обслуживания	Всего, чел
Алматинский	90	36	3	129
Актюбинский	60	21	3	84
Восточно-Казахстанский	68	12	2	82
Карагандинский	66	16	2	84
Южно-Казахстанский	68	16	4	88
Всего	352	101	14	467

III. Базовое техническое и программное обеспечение центров

6. В каждом из центров были организованы локальные вычислительные сети (ЛВС). Для их оптимальной работы создавались сегменты, состоящие из 1 сервера, 1 сканера, 1 персонального компьютера (ПК) для сканирования, 1 ПК для распознавания и 10 ПК для корректировки. Таким образом, всего в составе каждого из ЛВС было по 2 сервера, 2 сканера и 24 ПК. Все комплексы, специально созданные для обработки данных функционировали в сети Ethernet под управлением Windows NT Server.

7. Цветной сканер ScanStar 5045C позволяет сканировать 50 бланков в минуту, формата А4 при разрешимости 200 dpi (точек на дюйм) и плотности бумаги 70-80 г/м². В податочный лоток сканера одновременно загружается 100-150 бланков (теоретически допустимо 300), при этом в одной пачке могут быть перемешаны бланки разных типов. Практически сканировалось 27-30 бланков в минуту, т.к. качество бумаги не всегда соответствовало допустимой плотности, кроме того между порциями 1,000-1,200 бланков производилась чистка сканера от пыли. Ввод данных замедлялся также из-за невысокой скорости распознавания, которую можно было регулировать (чем медленнее шло распознавание, тем выше было его качество).

8. Цветные сканеры ScanStar 5045C и программные средства - BUSY, Image Port, JobScan, RecoStar являются продуктами фирмы CGK (Computer Gesellschaft Konstanz), являющейся дочерней структурой фирмы Siemens Nixdorf, Германия. Стоимость одного сканера со всеми базовыми программами составляла 67,770 долларов США. Общая стоимость проекта по поставке 11 комплектов оборудования, включая установку в кузовых центрах, организацию склада запасных частей и проведение обучения составила 995,680 долларов США.

9. Назначение базового программного обеспечения следующее: BUSY - программное обеспечение для автоматической обработки потока документов в ЛВС и разработки приложений, Image Port, JobScan - семейство программного обеспечения для управления процессом сканирования и для объединения сканера с системой обработки, RecoStar - программное обеспечение для распознавания знаков.

10. Скорость распознавания зависит от мощности компьютера, на котором установлена программа RecoStar и может достигать 100 символов в минуту. Ошибки распознавания составляют не более 1 процента при условии соблюдения необходимых требований к качеству бумаги и качественной печати бланков в типографии. Информационная насыщенность бланков составляла:

бланк 1Б - 22 символа, 1 метка,
бланк 2П – 334 символа, 14 меток,
бланк 3С – 141 символ, 21 метка,
бланк 4И – 142 символа, 6 меток.

11. Система BUSY прежде всего дает возможность пользователю управлять рабочим потоком документов и процессом обработки каждого из них, кроме того она предлагает средства по разработке приложений и по администрированию. Для разработки приложений система BUSY предлагает множество собственных функций, необходимых для контроля введенной информации, и кроме того предоставляет программисту возможность разрабатывать и подключать самостоятельно разработанные специфические процедуры, написанные на языке Visual C++. Программная оболочка BUSY способна контролировать все процессы обработки данных - сканирование переписных листов, распознавание, контроль, корректировку данных, преобразование и электронную архивацию данных.

12. Для разработки приложений были подготовлены специалисты, знающие не только новые системы Image Port, JobScan, Recostar, BUSY, но и получившие определенные навыки работы с пакетом Visual C++ и с сетевыми операционными системами. Специалисты из Германии провели обучение наших программистов работе со всеми программными продуктами, поступившими вместе со сканером. Кроме того, своими силами было проведено обучение пользователей из кустовых центров работе с разработанными комплексами программ. Были разработаны подробные инструкции для пользователей, где представлено полное описание всех этапов работ, описано как решать всевозможные проблемы, могущие возникнуть на каждом из этапов обработки.

IV. ЭТАПЫ ТЕХНОЛОГИЧЕСКОГО ПРОЦЕССА ОБРАБОТКИ ДАННЫХ

13. Выбранный нами технологический процесс обработки данных довольно сложен и трудоемок, но он позволяет добиться высокой скорости обработки, получить достаточно точные и качественные данные. Программы, основаны на алгоритмах, позволяющих выявить не только ошибки распознавания, но и ошибки счетчика, допущенные им при невнимательном и быстром заполнении бланков. Была обеспечена высокая степень автоматизации всех этапов обработки.

14. Весь процесс автоматизированной обработки материалов переписи населения состоит из двух уровней. В **кустовых центрах** решались следующие задачи:

- (a) Сканирование переписных листов;
- (b) Распознавание содержимого бланков;
- (c) Корректировка неверно распознанной или нераспознанной информации в среде BUSY, в три этапа - первичная, основная и с использованием имиджа (отсканированного образа);
- (d) Контроль на целостность данных - арифметический, логический и межбланочный;
- (e) Автоматическое кодирование, преобразование данных;
- (f) Загрузка преобразованных данных в СУБД Access (Система управления базой данных);
- (g) Корректировка в среде СУБД Access с использованием имиджа;
- (h) Контроль информации в разрезе района по основным показателям;
- (i) Архивирование, запись на CD-ROM или дискеты и передача в республиканский центр.

15. Схема технологического процесса обработки данных в одном сегменте ЛВС, образованного с включением одной системы сканирования и распознавания приведена в Приложении I.

16. На **республиканском уровне** выполнялись следующие задачи:

- (a) Прием информации с региональных центров;
- (b) Контроль информации на полноту и по основным показателям в разрезе районов (СУБД Access);
- (c) Создание базы данных по области (СУБД MS SQL Server);
- (d) Формирование таблиц по разделам;
- (e) Резервное копирование;
- (f) Анализ данных и их уточнение с областными управлениями;

- (g) Загрузка базы данных в разрезе областей в центральную базу;
- (h) Формирование сводных таблиц и генерация отчетов.

17. После ввода переписных листов осуществленного сканированием автоматически происходит распознавание введенных данных. Содержимое бланков подвергается специальным тестам. Результаты распознавания сравниваются с данными словарей, благодаря этому качество распознавания значительно улучшается. После распознавания пользователю предоставляется возможность корректировки переписных листов. Данные каждого портфеля записываются в отдельный файл и проходят по мере необходимости несколько стадий контроля и корректировки.

18. **Особенности организации процесса корректировки.** Разработана многоэтапная система контроля и корректировки, все данные проходят в первую очередь контроль на целостность данных, далее арифметический, логический и межтабличный контроли.

19. Корректировка разбивается на три последовательных этапа - первичная, основная корректировка и корректировка специалиста. На первичной корректировке исправляются все ошибки связанные с плохим распознаванием символов. Здесь выполняется проверка на наличие сопроводительного бланка или его дублирование. Далее проверяется код территории в бланке «Б» (если код территории правильный, то он автоматически проставляется в остальных бланках). Так же на первом этапе проверяются показатели, значения которых содержатся в словаре (место рождения, национальность, гражданство, государственный язык). Здесь же происходит проверка на дублирование номера бланка и на наличие всех бланков. В случае необходимости можно выполнить перенумерацию бланков блоками.

20. После первичной корректировки происходит переход на основную корректировку, где сразу начинает работать логический контроль, при этом программа анализирует ошибки и там, где возможно, ошибки исправляются программным путем (выполняется автоматическое проставление меток и значений согласно определенному алгоритму, разработанному методологами).

21. Если в процессе основной корректировки возникают проблемы, которые не могут быть разрешены простым корректировщиком, то происходит переход на следующий этап, где ошибки исправляются на уровне специалиста – эксперта по переписи населения.

22. Таким образом, можно отметить следующие методы оптимизации использованные при организации обработки данных. Во первых - это организация нескольких этапов корректировок, от простой механической, не требующей осмысливания результатов до этапа корректировки, где корректировщику требуется определенный уровень знаний и

специальная подготовленность. Во вторых - при корректировке пользователю предлагается альтернатива поправок, которые надо только выбрать и подтвердить. При этом рядом с распознанным текстом высвечивается соответствующий текст из словаря для проведения сравнения данных, или программа предлагает «всплывающие» списки таким образом, что пользователь может выбрать в списке нужный текст для корректировки. Средства системы BUSY и Recostar позволяют использовать словарную технологию при распознавании рукописных букв для автоматической корректировки данных. При распознавании происходит обращение к словарю, результаты распознавания текста сравниваются с соответствующими данными словаря и при необходимости автоматически корректируются. Здесь же результаты распознавания сравниваются с данными таблиц триграмм, содержащих наиболее часто используемые трехбуквенные комбинации и после сравнения автоматически корректируются. После того как распознанный текст прошел контроль, происходит автоматическое кодирование данных, вместо текста в файл вносится числовое значение кода.

23. После завершения корректировки данные автоматически кодируются. Отконтроллированные и очищенные данные каждого портфеля выгружаются в базу данных СУБД MS Access, где производится очередной контроль и при необходимости данные опять подвергаются корректировке. Здесь же могут быть получены предварительные результаты по основным показателям в пределах одного района и если возникает потребность, то из этого этапа возможен возврат обработанного портфеля на повторную корректировку в систему BUSY, где до окончательного завершения корректировки хранятся имиджи бланков возвращаемого портфеля. Из СУБД MS Access полностью проверенные данные поступают в СУБД MS SQL Server, где накапливается информация и создается копия базы данных. СУБД MS SQL Server является одной из наиболее мощных серверов баз данных и позволяет работать в архитектуре «клиент-сервер». При такой архитектуре клиентское приложение формирует запрос к серверу базы данных, на котором выполняются все команды. Результаты команд посылаются клиенту для использования, просмотра и печати.

24. Данные, которые сформировались в территориальных подразделениях, выгружаются для передачи в центр по каналам связи, а также на компакт-дисках. На республиканском уровне создана централизованная база данных переписи. Он функционирует на основе СУБД MS SQL Server. Некоторые сведения об объемах баз данных приведены в Приложении II.

V. ПРИКЛАДНЫЕ ПРОГРАММЫ, РАЗРАБОТАННЫЕ ДЛЯ ОБЕСПЕЧЕНИЯ ЭТАПОВ ОБРАБОТКИ

25. Внедрение новых и достаточно сложных технологий, использование множества этапов при обработки материалов переписи потребовали применения большого количества новых программных продуктов. Возникла необходимость в разработке многих комплексов приложений, а именно:

- (a) Комплекс приложений с применением систем Image Port , JobScan, Recostar, BUSY:
 - (i) Создание описаний документов;
 - (ii) Разработка системы контроля введенной информации;
 - (iii) Разработка программ корректировки данных;
 - (iv) Разработка программ преобразования данных и занесения информации из системы Busy в итоговую базу данных.
- (b) Комплекс приложений с применением Visual C++, Access, Ms SQL:
 - (i) Программы контроля , корректировки данных , выгруженных из системы Busy в Access;
 - (ii) Разработка вспомогательных программ обработки материалов переписи;
- (c) Комплекс приложений, предназначенных для общего управления потоками информации в системе:
 - (i) Программы преобразования данных;
 - (ii) Программы формирования и ведения нормативно-справочной информации;
 - (iii) Программы организация процессов учета и передачи информации из филиалов в центр.
- (d) Комплекс программ формирования регламентных таблиц:
 - (i) Разработка программ формирования технологических файлов с агрегированной переписной информацией.
 - (ii) Программы генерации отчетов.

26. Большое количество разработанных прикладных программ является характеристикой уровня сложности проекта. С другой стороны, это показатель высокого уровня специалистов-программистов, сумевших за короткий срок освоить и внедрить

достаточно сложные на тот момент информационные технологии. Именно здесь была впервые применена технология «клиент – сервер», и это позволило на практике убедиться в преимуществе такой схемы обработки большого потока данных.

VI. ОСНОВНЫЕ ПРОБЛЕМЫ, ВОЗНИКШИЕ ПРИ ОБРАБОТКЕ МАТЕРИАЛОВ ПЕРЕПИСИ НАСЕЛЕНИЯ

27. Погрешность при распознавании сильно зависела от качества переписных бланков, они были отпечатаны некачественно.

28. Слишком сжатые сроки повлияли на качество разработанных программ. Обучение специалистов было завершено только за 5 месяцев до начала переписи.

29. Не было возможности изучить опыт использования аналогичного оборудования другими организациями в переписи населения, т.к. подобный опыт в странах СНГ отсутствовал.

30. Пилотная перепись была проведена с использованием других бланков, предусматривалась другая технология. Из-за этого не был создан прототип будущей системы, не было проведено апробации.

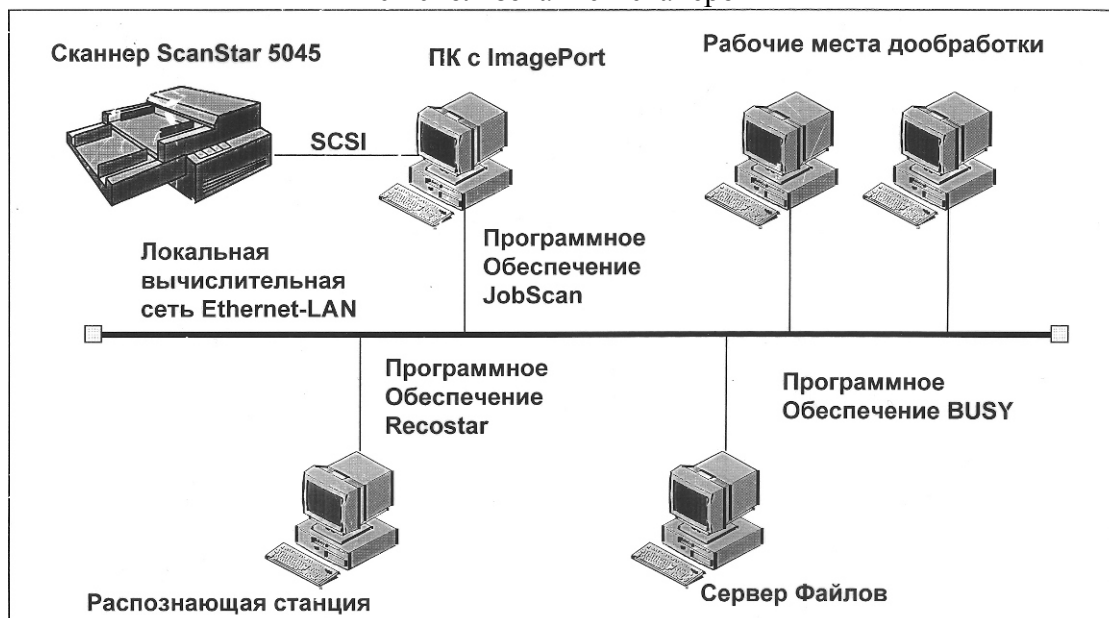
31. Базовое программное обеспечение, поставляемое вместе со сканером также имеет свои недостатки. Программа применима для узкого спектра функциональных требований – внедрение в коммерческих банках, не было опыта использования в переписи. Программа имеет встроенную функциональность, но она сильно чувствительна к внедрению дополнительных приложений, разработанных нашими программистами (всплывающие окна, система идентификации портфелей и т.д.), которые вызывают частые сбои и замедление работы системы. Не предусмотрен контроль внутри пачки документов (ранее на этой системе не был реализован межбланочный контроль).

VII. ЗАКЛЮЧЕНИЕ

32. Несмотря на наличие проблем и сложности организации ввод данных из переписных бланков завершился раньше планировавшегося срока. Все участники проекта впервые на практике увидели и оценили преимущества обработки материалов переписи с использованием сканеров. Опыт, полученный в ходе реализации этого проекта был применен в организации обработки материалов сельскохозяйственной переписи, которая проводилась в два этапа в 2006-2007 годах.

Приложение I

Схема технологического процесса обработки данных с использованием сканеров



1. Для осуществления работ по сканированию, сканер ScanStar 5045C соединен с ПК, на котором установлены программы Image Port, JobScan.
2. После сканирования пакет документов (файл) со сканирующего ПК отправляется на сервер файлов, на котором установлена серверная часть программы BUSY.
3. В автоматическом (или ручном) режиме файл с сервера перебрасывается в ПК «Распознающая станция», на котором установлено программа RecoStar.
4. После распознавания файл автоматический возвращается на сервер.
5. На всех ПК группы «Рабочие места дообработки» с установленными клиентскими частями программы BUSY осуществляется процесс корректировки.
6. На одном из ПК группы «Рабочие места» устанавливается СУБД ACCESS и осуществляется корректировка и контроль на полноту в разрезе районов.
7. При выполнении пунктов 5 и 6 на помощь оператору на экран ПК выдается имидж переписного документа.
8. Данные по районам после контроля заносятся в базу данных по области, которая функционирует в СУБД MS SQL Server.

Приложение II

Таблица. Сведения об объеме базы данных переписи населения

	Области	Количество записей в базе данных	
		по бланку 3С	по бланку 2П
1	Акмолинская	843,245	832,267
2	Актюбинская	690,859	679,438
3	Алматинская	1,569,571	1,533,758
4	Атырауская	450,694	437,857
5	Восточно-Казахстанская	1,539,184	1,524,635
6	Жамбылская	998,445	980,310
7	Западно-Казахстанская	623,931	610,052
8	Карагандинская	1,422,851	1,389,921
9	Костанайская	1,024,333	1,001,850
10	Кызылординская	602,248	593,315
11	Мангистауская	319,442	312,998
12	Павлодарская	809,883	803,257
13	Северо-Казахстанская	730,056	722,280
14	Южно-Казахстанская	1,990,444	1,972,187
15	Астана	324,758	318,769
16	Алматы	1,156,806	1,096,483
	Республика Казахстан	15,096,750	14,809,377
	Объем базы данных, Гб	4,5	1,62

В таблице приведены сведения по объему базы данных, сформированной после первичной обработки бланков – считывания и корректировки. Объем базы данных сформированной после окончательной обработки составляет 13 Гб. Каждая запись в базе соответствует одному физическому лицу и состоит из 66 полей, имеется возможность выдачи данных в разрезе населенного пункта, района, области. Информационная система «Перепись населения» функционирует в СУБД MS SQL Server, ее составными частями являются следующие объекты:

- а) базы данных 16 областей (основной ресурс);
- б) база нормативно-справочной информации (классификаторы, словари, справочники);
- с) модули формирования выходных форм.
