Государственный комитет
статистики Украины

Опыт использования сканеров при обработке данных переписи населения 2001 года

1. Обработка данных Первой национальной переписи населения Украины 2001 года была сложным, но в то же время успешным проектом. Анализ всего комплекса работ, реализованных во время обработки материалов первой переписи населения и анализ ошибок очень важен именно сейчас, когда началась подготовка к проведению следующей переписи населения.

2. По опыту предыдущих переписей наиболее напряженным этапом всегда был этап ввода, контроля и корректировки информации переписных листов (ПЛ). Именно на этом этапе формируется информационная база для всех дальнейших расчетов. Своевременное и качественное выполнение этапа формирования информационной базы переписных листов определяет успех всей переписи в целом. Характерной чертой этого этапа является взаимодействие значительного количества людей, которые должны были в довольно сжатые сроки обработать большие объемы (в целом по Украине - около 70 млн.,) переписных листов. Все остальные этапы обработки материалов переписи в большей степени зависят от имеющихся технических и программных средств, т.е. в большей степени контролируемы. На наш взгляд, этап формирования баз данных на региональном уровне является ключом к успеху машинной обработки данных. Поэтому при выборе технологии обработки мы значительное внимание уделили этапу ввода данных.

Следующим характерным моментом для обработки материалов переписи является стопроцентная загрузка персонала и технических средств на протяжении всего периода обработки.

Учитывая все это, а также значительные объемы бумажных носителей, было принято решение о децентрализованном формировании информационного фонда. При этом ввод данных переписных листов и уточнение их содержимого велось на региональном уровне. С целью обеспечения сроков выполнения работ по обработке материалов переписи в каждом региональном управлении статистики был создан специализированный технологический участок. Откорректированные данные в виде территориальных файлов на машинных носителях передавались на центральный уровень в Главный вычислительный центр Госкомстата Украины.

3. Технология обработки Всеукраинской переписи население 2001 года использовала сертифицированные технические и программные средства, которые отвечают требованиям открытых систем взаимодействия. Технология строится с использованием принципов "клиент-сервер".*Первым этапом работ по обработке информации было введение данных с бумажных носителей устройствами оптического считывания с проведением контроля (в ограниченном объеме), подготовка файлов переписных листов для загрузки в базу данных; формирование архива графических изображений*. Этот этап выполнялся с помощью программно-

технического комплекса сканерного ввода переписных документов, который был построен на основе программного обеспечения Eyes&Hands Forms шведской компании ReadSoft.

Для оптического считывания информации с переписных бланков использовался сканер **Fujitsu M4099GH, который позволяет сканировать 50-55 двусторонних бланков в минуту.** Формат бумажных носителей - А4, при разрешимости 200 dpi (точек на дюйм) и плотности бумаги 70-80 г/м. Емкость подающего кармана – до 1000 бланков, при этом в одной пачке могут быть бланки разных типов. Кроме того между порциями 20 тис. бланков производилась чистка сканера от пыли, капитальная профилактика проводилась после 100 тис.

Стоимость одного сканера составляла 25 тис. долларов США. Общая стоимость проекта по поставке 30 комплектов оборудования, включая установку в областных управлениях, закупка запасных частей, проведение обучения персонала и разработку необходимых программных средств, составила около 6 млн. долларов США.

Этот комплекс был предназначен для:

- сканерного ввода переписных документов;
- автоматического (компьютерного) распознавания переписных документов;
- проверки автоматического распознавания и обеспечения первичного контроля данных переписных документов;
- формирования текстового файла с информацией, которая подлежит обработке на дальнейших этапах;
- записи графических образов переписных документов, полученных в результате сканирования

5. Комплекс был реализован по модульной технологии с поэтапным выполнением работ. Это позволило организовать выполнение этапов процесса сканирования на нескольких компьютерах с распределением его во времени. Количество одновременно запущенных модулей Eyes&Hands Forms должно было превышать восьми.

6. В состав программного комплекса сканирования входило 5 модулей Eyes&Hands Forms: Manager, Scan, Interpret, Verify, Transfer и отдельный модуль записи графических образов CensusCDR.

Название модуля	Назначение модуля
Manager (Управление)	Описание переписных форм. Удаление неудачно обработанных переписных документов и очищение внутренней базы переписных документов
Scan (Сканирование)	Сканирование переписных документов портфеля
Interpret (Распознавание)	Распознавание просканированных переписных документов
Verify (Верификация Проверки)	Проверка оператором правильности распознавания
Transfer (Трансформации)	Формирование исходного файла и каталога графических образов переписных документов по проверенному портфелю
CensusCDR (Записи дисков)	Запись графических образов переписных документов на лазерный диск

7. Работу с комплексом осуществляли две категории операторов: сканирования и верификации. Оператор сканирования выполнял сканирование переписных документов, запуск модулей Interpret и Transfer, запись графических образов переписных документов. Оператор верификации - запуск модуля Verify и проверку правильности распознавания цифровых символов переписных документов с помощью модуля Verify.

8. **Все другие этапы** обработки выполнялись с использованием программного обеспечения АС “Перепись-2001”, который был разработан сторонней компанией "Квазар-Микро Техно":

- **формирование отчетной документации о ходе обработки материалов переписи; регистрация и возвращение обработанных портфелей переписных листов в хранилища**.
- **загрузка файлов переписных листов в базу данных.** Подготовленные на этапе оптического считывания файлы загружались в базу данных. Единицей загрузки был портфель.
- **контроль переписных листов, корректировка данных переписных листов на основе протоколов контроля с использованием архива графических изображений; выполнение повторного контроля.**
- **формирование сводных данных в разрезе массивов, проведение внутритабличного, межразрезного и межтабличного контролей.**
- **формирование территориального файла (части территориального файла) для отсылки (передачи) на государственный уровень; создание архивных копий**.
- **получение от территориального управления статистики, регистрация и антивирусный контроль территориального файла на государственном уровне; подготовка территориального файла для загрузки в центральную базу данных**.
- **загрузка территориального файла в центральную базу данных переписи населения**.
- **выполнение контроля переписных листов, которые входят в территориальный файлов; возможно, повторные запросы территориальных файлов или их частей**.
- **формирование сводных данных (выходных таблиц), проведение их внутритабличного, межразрезного и межтабличного контроля**.
- **печать выходных таблиц**.
- **формирование информации для пользователей данных**.
- **формирование региональных фрагментов центральной базы данных переписи населения и отсылка (передача) их в территориальные управления статистики для дальнейшего использования; печать сводных данных (выходных таблиц) в территориальных управлениях статистики**.
- **подготовка материалов Всеукраинской переписи населения для распространения печатными средствами, средствами Internet и формирование файлов на магнитных и оптических дисках.**

9. Программно - технологический комплекс (ПТК) обработки данных переписи населения АС "Перепись - 2001" состоит из регионального уровня и государственного. На региональном уровне от должен был обеспечивать:

- загрузку в базу данных информации, полученной путем сканерного ввода и формирование протокола загрузки;
- создание архива графических изображений и получение данных из него в пределах внешнего интерфейса для обработки переписных листов;
- удобную систему навигации по данным переписи населения;
- просмотр и печать нормативно - справочной информации;
- интерактивное введение и корректировку переписных листов;
- сплошной и фрагментарный контроль переписных листов с формированием протокола;
- автокорректирование переписных листов с формированием протокола;
- формирование протоколов о составе базы данных, качестве данных, состоянии прохождения обработки материалов переписи населения, и другой информации для принятия организационных решений во время обработки данных;
- формирование сводных данных (выходных таблиц);
- проведение **внутритабличного, межразрезного и межтабличного** контролей выходных таблиц с формированием протокола;
- полную и частичную печать протоколов контроля;
- систему поддержки нерегламентированных запросов к базе данных;
- систему копирования и восстановления материалов переписи населения, ведение и обновление архивных копий;
- формирование территориальных файлов для передачи на государственный уровень;
- отсылку откорректированной информации (фрагментов территориальных файлов) на государственный уровень имеющимися средствами связи;
- загрузку региональных фрагментов центральной базы данных переписи населения;
- печать выходных таблиц в необходимых разрезах и их вывод на технические носители;
- систему поддержки администрирования и размежевания прав доступа к данным переписи населения.

10. Для государственного уровня ПТК АС "Перепись - 2001"должен был выполнять тот же перечень функций, что и регионального уровня, и дополнительно обеспечивать:

- введение и корректирование нормативно-справочной информации;
- загрузку территориальных файлов в центральную базу данных переписи населения;
- выгрузку региональных фрагментов центральной базы данных переписи населения;
- формирование информации переписи населения для пользователей данных.

11. В процессе эксплуатации АС ”Перепись-2001” часто изменяются (накапливаются) как ее функциональные возможности, так и обрабатываемые данные. Объем данных, накопленных в процессе работы с системой, имеет колоссальный размер, но для анализа и изучения доступны лишь те данные, которые предусмотрены функциональными возможностями системы (фиксированный перечень как выходных таблиц, так и разрезов, по которыми они строятся).

12. Поэтому после разработки и внедрения АС ”Перепись-2001” возникла необходимость в создании на основе применения OLAP-Технологии новой системы - АС ”Перепись-2001 Аналитик”.

13. Основным назначением системы АС ”Перепись-2001 Аналитик” является предоставление возможности проведения анализа данных консолидированной базы данных АС ”Перепись-2001” широкому кругу специалистов по демографии, переписи населения и других областей статистики (и не только статистики).

14. Системой обеспечивалась реализация следующих задач:

- построение выходных таблиц по произвольно сформированным разрезам;
- предоставление удобного визуального интерфейса управления процессом формирования данных;
- предоставление удобного образно-визуального интерфейса отображения сформированных данных;
- минимизация затрат времени при получении статистических данных;
- облегчение анализа данных благодаря оперативности их получения и образности отображения.

Таким образом, пользователь получает естественную, интуитивно понятную модель данных, организовывая их в виде многомерных кубов. Измерениями куба выступают такие характеристики данных, в разрезе которых можно получить, отфильтровать, сгруппировать и отобразить информацию.

Для построения запросов к многомерному кубу и визуализации полученных результатов используется подсистема построения гибких запросов, построенная на основе технологии Microsoft Excel PivotTable. Полученная аналитическая информация может быть представлена либо в виде динамической выходной таблицы произвольного разреза, либо в виде разнообразных диаграмм, графиков, гистограмм, которые визуально демонстрируют закономерности распределения данных в зависимости от определенных выбранных разрезов.

Проблемы:

17. Пробная перепись была проведена с использованием других бланков, неудачный выбор цвета зон считывания на переписных доекументах-носителях – желтый. Этот цвет очень негативно воспринимался глазами на протяжении

долгой кропотливой роботы с документами и был одной из причин быстрого утомления кодировщиков.

18. Качество заполнения переписных листов (текучесть кадров: счетчиков, кодировщиков. Не хватало времени провести более качественное их обучение)

19. Т**олько в июле 2002** года (планировалось с апреля) региональные управления начали сканирование бланков (считывание и проведение контроля введенных данных, предполагалось проводить параллельно с кодированием) из-за разработки соответствующего программного обеспечения с задержкой на четыре месяца. Если бы сканирование проводилось параллельно с кодированием данных, то ошибки, которые были выявлены при помощи специальных контролей после введения первых порций информации, можно было бы легко избежать и исправить. Но большая часть массива ПЛ к началу сканирования уже была закодирована.

20. Отсутствие опыта работы у специалистов регионального и центрального уровней с таким объемом информации.

21. Низкое качество печати переписных листов (отдельные точки, черточки, загрязнения в зонах считывания, смещение линий и т.д.)

22. Но несмотря на трудности все региональные управления **уже в ноябре 2002** года с успехом справились с поставленной задачей и закончили ввод переписной документации, ее запись на технические носители, формальный и логический контроль данных.

23. К середине декабря 2002 г. в каждом регионе был сформирован территориальный файл и передан на ГМУС для формирования основных выходных таблиц по Украине и регионам для официальной публикации данных ВПН. Первые итоги (о численности и составе населения Украины) были опубликованы ГКС 25 декабря 2002 г. По регионам данные об общей численности населения, его поло-возрастной структуре, национальной и языковой структуре, распределению по гражданству, образовательным уровням и семейному состоянию были предоставлены общественности в январе 2003 года.

24. Таким образом выбранный нами технологический процесс обработки данных довольно сложен и трудоемок, но он позволил 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.