Distr.
GENERAL

CES/SEM.40/6
18 September 1998

RUSSIAN
Original: ENGLISH

СТАТИСТИЧЕСКАЯ КОМИССИЯ и ЕВРОПЕЙСКАЯ ЭКОНОМИЧЕСКАЯ КОМИССИЯ

СТАТИСТИЧЕСКОЕ УПРАВЛЕНИЕ ЕВРОПЕЙСКИХ СООБЩЕСТВ (ЕВРОСТАТ)

КОНФЕРЕНЦИЯ ЕВРОПЕЙСКИХ СТАТИСТИКОВ

Совместная рабочая сессия ЕЭК-Евростата
по переписям населения и жилищ 1/
(Дублин, Ирландия, 9-11 ноября 1998 года)

Тема 2

ОБРАБОТКА ИЗОБРАЖЕНИЙ В ФОРМАТЕ ASCII

Обработка данных переписи населения и жилищ 1995 года в Израиле

Специальный документ, представленный Центральным статистическим бюро Израиля 2/

1. Сбор первичных данных и преобразование их в значимую статистическую информацию являются основными производственными процессами статистических органов. Эти процессы претерпевают различные изменения и усовершенствования, целями которых, в частности, являются их адаптация к изменяющимся потребностям пользователей и разработка статистических методов, позволяющих более объективно отражать происходящие явления.

1/ Обработка документации для этой рабочей сессии будет вестись на уровне семинара.

2/ Авторы: Оливия Блюм и Элиаху Бен-Моше.

Однако в нынешнюю информационную эпоху все более высокие требования, предъявляемые к статистическим органам с точки зрения представления надежной, своевременной и обновленной информации, также могут стимулировать технологические развитие, изменения и адаптационные меры.

2. Ведущее место в этих процессах всегда отводилось переписям с учетом их значимости для национальной статистики, масштабов и, следовательно, затрат по проведению.

3. В начале 90-х годов методика проведения переписей обогатилась разработками в области "оконной" технологии и усовершенствованных* методов оптического считывания. Цели переписи были не изменены, а сформулированы по-новому с учетом нововведений в методах регистрации и обработки данных. Возможность открывать одновременно несколько окон, обладающих оптическими и физическими взаимосвязями, а также наличие более надежной технологии оптического распознавания символов заставили нас по-иному подойти к формулировке задач в области обработки данных.

4. В ходе переписи населения Израиля 1995 года использовался новый подход к обработке результатов переписи, опиравшийся на преимущества новых технологий, существовавших на тот момент, и накопленный до этого времени международный опыт.

5. В настоящем документе рассматривается ряд аспектов воздействия меняющейся технологической среды на основные составляющие элементы переписи. В настоящее время мы занимаемся усовершенствованием идеологии, лежащей в основе планирования переписи, обработки данных и контроля качества процессов.

Планирование операций с учетом возможностей системы оптического ввода данных (ОВД)

6. Изменение технологической рабочей среды может осуществляться следующими двумя основными путями: внедрение новых средств для осуществления одних и тех же задач и процессов (как это делалось ранее) или оптимизация использования этих новых ресурсов при одновременном изменении логических принципов системы и ее компонентов.

---

* Технология оптического распознавания была разработана несколько десятилетий назад и используется для проведения полномасштабных и пробных переписей с начала 70-х годов.

A: Первая стратегия, заключающаяся в обычной компьютеризации процесса ввода данных, позволяет сократить объем времени, необходимого для осуществления процесса и подготовки высококачественных массивов данных переписи за счет снижения влияния субъективного человеческого компонента. Задачи по вводу данных передаются от людей машине, вследствие чего процессы приобретают единообразный и непротиворечивый характер. Конечный результат в данном случае обладает большей степенью надежности по сравнению с менее компьютеризированными альтернативами. Применение компьютеризированной системы также может вести к снижению затрат, если ее внедрение не требует технологических разработок.

B: В то же время вторая стратегия, предусматривающая наряду с внедрением новых технологий пересмотр идеологии, лежащей в основе всего процесса переписи, обладает дополнительными преимуществами, выражающимися в усовершенствовании организационных, а также операционных аспектов:

B1: Организационные аспекты

7. В технологической среде последовательность этапов не определяется логистикой обработки бумажных опросных листов и, следовательно, может изменяться, игнорироваться или опираться на новые логические принципы системы. Примером этому может служить отсутствие необходимости выбора того, что делать в первую очередь - редактировать данные или кодировать их. Копии всех изображений могут быть одновременно направлены на оба этапа обработки. Таким образом, обе задачи могут осуществляться одновременно, если только логический принцип, например, проведение редактирования уже закодированных данных, требует иного подхода. Аналогичным образом в связи с созданием новой рабочей среды возникает вопрос о том, следует ли осуществлять ввод отредактированных и закодированных данных или сначала вводить первичные данные, а затем уже осуществлять эти процедуры обработки. Вопрос последовательности и логистики приобретает еще более важное значение, если он не ограничивается рамками процесса ввода данных, а также учитывает процессы, осуществляемые до или после этого этапа. Сбор опросных листов в определенном порядке теряет свое значение, если на каждой странице содержатся поддающиеся идентификации переменные. Они могут загружаться в блок питания сканнера в любом порядке или в любом желаемом порядке с учетом статистических соображений. Кроме того, идентификация дублирующих записей, которая обычно осуществляется центральным компьютером на этапе макроредактирования, может производиться в ходе процесса ввода данных по мере поступления опросных листов.

B2: Операционные аспекты

8. Операционные аспекты оптимизации использования ресурсов компьютеризированной системы оптического распознавания охватывают все составляющие элементы переписи. Структура опросных листов должна быть разработана с учетом требований оптического считывания. Они должны содержать поддающиеся идентификации переменные по каждой

физической и логической единице, причем для облегчения оптического распознавания предпочтение должно отдаваться вопросам, допускающим единственный ответ и т.д.

9. Помещение компьютера в начало процесса ввода данных открывает возможность пересмотра содержания всех последующих задач; ввод с клавиатуры становится подтверждением оптического распознавания, микроредактирование может быть сведено до минимума или даже полностью исключено, а кодирование осуществляться с помощью интерактивных запросов и т.д.

10. Кроме того, если файл используется для условных расчетов отсутствующих величин методом подстановки, осуществляемых на стадии макроредактирования в центральном компьютере, он может быть увязан с результатами переписи сразу же после завершения первого этапа оптического распознавания. Поскольку этот файл не должен копироваться в пассивный архив до получения всех данных переписи, этот внешний файл может являться частью процесса ввода данных, подтверждать оптическое распознавание символов и использоваться для идентификации дублирующих записей, поскольку изображения соответствующих опросных листов будут по-прежнему храниться в системе. Он также может использоваться для условного расчета переменных, если микроредактирование осуществляется на этапе ввода данных.

11. Хотя цели переписи остаются одними и теми же независимо от уровня технологической оснащенности рабочей среды, использование новых средств содействует повышению гибкости планирования рабочих процессов и достижению новых промежуточных целей. В системе, основанной на технологии оптического распознавания символов, возможность доступа к изображениям опросных листов в ходе и после завершения этапа ввода устраняет необходимость проведения редактирования и кодирования в начале процесса, поскольку опросные листы находятся перед оператором. Таким образом, в ходе планирования создание файла первичных необработанных данных может рассматриваться в качестве промежуточной цели.

12. В рамках предыдущих переписей не ставилась задача по формированию файла необработанных данных и поэтому он никогда не создавался. В силу этого невозможно было найти после завершения этапа ввода информацию, которая содержалась непосредственно в опросных листах. Также отсутствовала возможность провести дифференциацию ошибок, внесенных в опросные листы респондентами или счетчиками, ошибок, введенных в ходе ручного редактирования или ошибок, внесенных на этапе ввода данных или же на других этапах. Постановка задачи по созданию файла необработанных данных ведет к изменению логических принципов, а также графика осуществления различных этапов обработки данных.

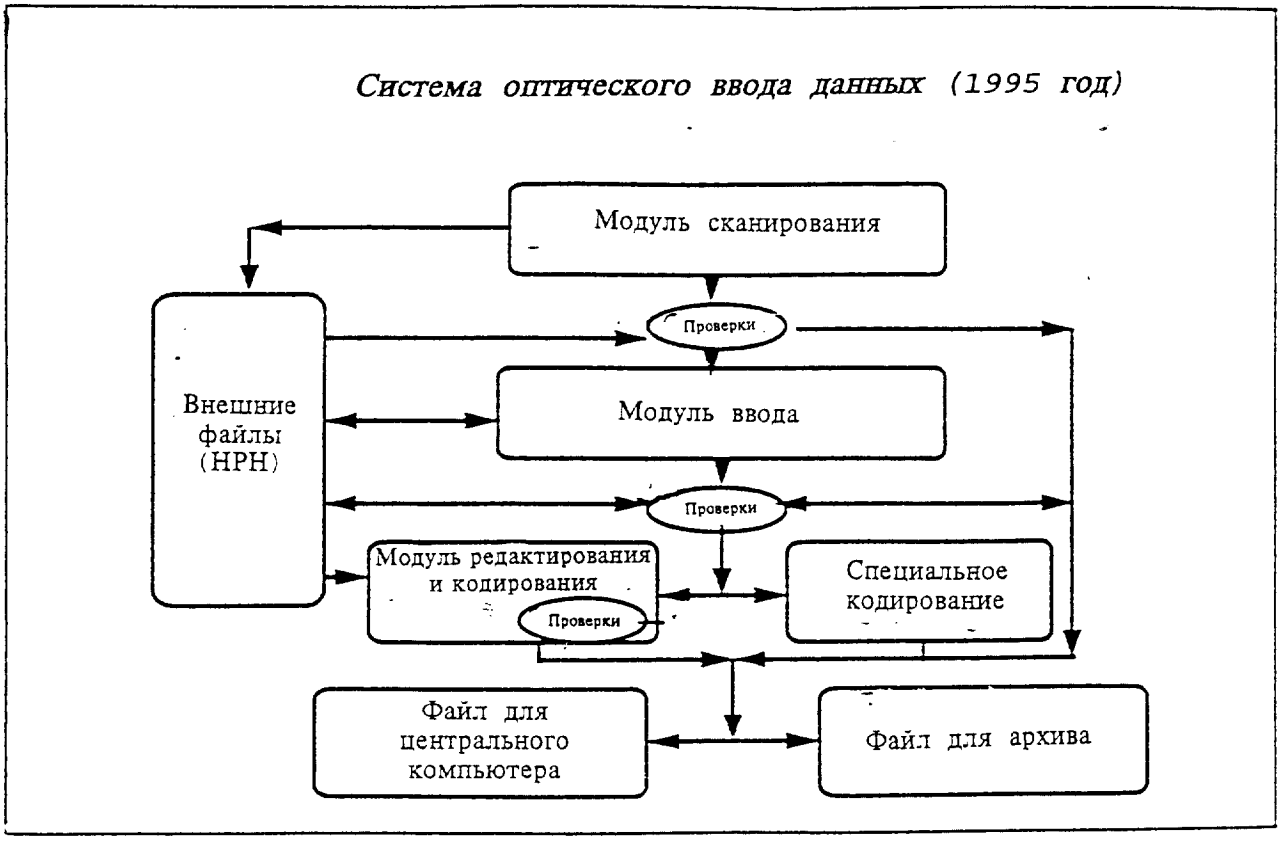
13. Планирование переписи населения и жилищ Израиля 1995 года проводилось на основе стратегии общей оптимизации. Усовершенствованная технология оптического распознавания символов и рабочая среда WINDOWS являлись исходными предпосылками для пересмотра в ходе планирования переписи содержания процессов и промежуточных целей, а также внедрения новых технологических усовершенствований.

Нижеследующие разделы посвящены рассмотрению трех вопросов:

1) обработка данных в вышеописанной рабочей среде;
2) использование внешних файлов в компьютеризированной системе;
3) опросный лист и система ОВД.

Обработка данных с использованием системы оптического распознавания

14. Система обработки данных состоит из следующих основных модулей: сбор данных, ввод данных, кодирование, редактирование, анализ данных и контроль качества в ходе процесса обработки. Система ОВД, использовавшаяся в рамках переписи населения и жилищ Израиля 1995 года, не имела модулей сбора и анализа данных. Ее структура выглядела следующим образом:



15. Контроль качества является составной частью системы. Он выполняет функции основополагающего параметра всех модулей, а также связывающего механизма между ними. Способ осуществления каждой из этих операций зависит от двух основных факторов, а именно - целей и средств.

16. Целью ввода данных переписи является создание файла результатов переписи с известными погрешностями. При наличии технологически усовершенствованной системы, обеспечивающей доступ к изображениям опросных листов на протяжении всего процесса, цель может заключаться в создании структурного файла необработанных данных, отражающих максимально по-возможности точно ответы респондентов. Этот файл может использоваться для локализации ошибок по их источнику и возвращения к исходным данным в случае необходимости пересмотра одного или более процессов. Возможность создания нескольких файлов, прошедших различные процедуры редактирования, позволяет производить оценку процедур редактирования с использованием компаративных методов и целенаправленное редактирование файла необработанных данных с учетом различных потребностей. Данные файлы используются не только для разработки общих материалов переписи, но и для принятия решений в отношении методики редактирования крупных файлов данных в будущем.

17. Все процедуры, осуществляемые в ходе процесса ввода данных, были разработаны с учетом поставленной цели, которая заключалась в оптимизации использования ресурсов существующей рабочей среды.

Руководящие принципы каждого модуля

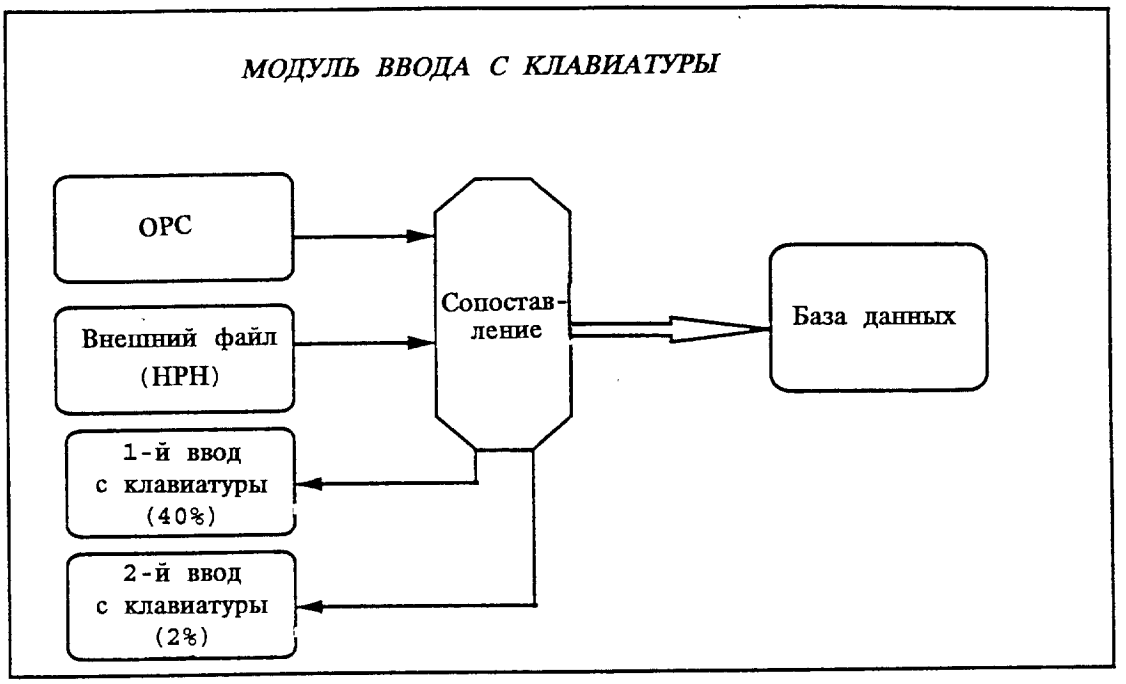
18. Основными задачами, выполняемыми **модулем сканирования**, являются проверки на полноту и исключительность, а также оптическое распознавание меток (ОРМ) и оптическое распознавание символов (ОРС).

19. Целью проверок на полноту является обеспечение сканирования обеих сторон каждого листа и ожидаемого числа опросных листов. Целью проверок исключительности является обеспечение только одноразового сканирования каждого раздела опросного листа.

20. Целью процессов ОРМ и ОРС является присвоение величин каждому полю, определенному в качестве цели для оптической идентификации, и статуса надежности оптического распознавания. Эти статусы (сверхнадежный, надежный, сомнительный и неудовлетворительный) определяют дальнейшую обработку символа или поля в системе.

21. Для эффективного использования вычислительных ресурсов в модуль сканирования изображение каждого опросного листа вводится только один раз. Фиксированный макет опросных листов (маска) выводится из системы по окончании сканирования. Кроме того, идентифицированные величины и присвоенные им статусы подвергаются сжатию перед вводом их в базу данных.

22. Задачами **модуля ввода с клавиатуры системы ОВД** является дополнение и проверка оптически считанных величин. Величина того или иного поля рассматривается в качестве прошедшей проверку, если два источника идентификации указывают одну и ту же величину. Ввиду того, что национальный регистр населения (НРН) увязывается с файлом сразу же после сканирования, а также наличия двух этапов ввода сопоставление величин осуществляется по следующей схеме:



23. После проведения сопоставления между ОРС (включая ОРМ) и НРН идентичные единицы направляются в базу данных. Остающиеся величины подвергаются первому этапу ввода с клавиатуры, после чего проводится сопоставление между ОРС, НРН и результатами первого ввода. Если два из трех результатов являются идентичными, величина направляется в базу данных. Остающиеся величины проходят этап повторного ввода с клавиатуры, после чего повторяется процесс сопоставления. При этой последовательности только 40% символов (за исключением буквенных) подвергаются первому вводу и лишь 2% - повторному вводу с клавиатуры (см. также раздел "Использование внешних файлов"). Громадная экономия трудозатрат операторов на этапе ввода не ограничивается только вышеописанным автоматическим процессом; ввод данных с клавиатуры представляет сам по себе "интеллектуальный процесс", в котором используется специально разработанная технология ввода изображений. Существует три типа (уровня) ввода:

- проверочный ввод с помощью "карт" - оператор работает с изображениями размером 10x10 символов, идентифицированных в качестве одной и той же цифры или метки, на которых он должен указать ошибки. Так, например, на карте цифры "8" помечаются все символы, не являющиеся "8". Настройка этой системы позволяет максимальную погрешность в размере 4-5 неправильных символов. Это означает, что 4-5 щелчков мышью позволяют проверить 95-96 символов, которые не направляются для дополнительной обработки в данном модуле.

- Корректировочный ввод последовательностей из трех символов - оператор работает с изображениями символов, полученных с различных опросных листов, и вводит их тройками. Решение об организации их в тройки основывается на эксперименте, а результаты, представляющие из себя тройки символов, являются оптимальным числом для запоминания и ввода с относительно высокой скоростью.

- Ввод полных полей - интерфейс представляет собой изображение одной и той же переменной в различных опросных листах. Так, например, изображения 12 полей ответов на вопрос "страна рождения" должны вводиться полностью. Формирование однородных элементов ввода, как показывает практика, содействуют повышению скорости и надежности ввода данных.

24. Уровень ввода с клавиатуры определяется уровнем надежности распознавания ОРС, введенной величиной (соответствует она или нет допустимому интервалу значений) и подтверждающей вспомогательной информацией. Исходно статус "сверхнадежный" означает, что величина была подтверждена и не требует повторного ввода, "надежные" величины предназначены для проверки в "картах", "сомнительные" - для ввода тройками, а "удовлетворительные" - для ввода полными полями.

25. Контроль качества является неотъемлемой частью процесса и основывается на использовании синтетических символов, включенных во вводимые элементы. Он позволяет осуществлять непрерывный контроль качества ввода с клавиатуры и качества работы операторов.

26. Задача по созданию структурированного файла необработанных данных требует осуществления следующих четырех основных процедур **редактирования**: определение структурных единиц, модификация и дополнение результатов ввода клавиатуры в тех случаях, когда конечная величина не определена, подтверждение или модификация введенных величин в полях, в которых были найдены логические противоречия (без изменения записанных в опросном листе величин), кодирование величин по категории "прочие" (открытая категория в вопросе, предусматривающем единственный ответ).

27. Для осуществления этих процедур в ходе или сразу же после завершения этапа ввода с клавиатуры используется набор подготовительных задач. Они включают в себя автоматическое кодирование двух переменных ("страна рождения" и "отношение к

основному лицу"), обнаружение письменных ответов в категории "прочие". При этом все физические (переписной участок, опросный лист) и логические (домохозяйства, частные лица) единицы автоматически определяются и проходят проверку на соответствие допустимому интервалу значений и на непротиворечивость. Выполнение одной из подготовительных задач создает элемент редактирования, в котором идентифицированы все проблемы редактирования, найденные в ответах одного домохозяйства.

28. Рабочая среда редактирования позволяет одновременно визуализировать изображение полной страницы опросного листа, изображения всех страниц, относящихся к домохозяйству, а также изображения ответов любого домохозяйства, входящего в состав переписного участка. Также можно одновременно визуализировать письменные ответы на опросный лист, внесенные счетчиком и респондентами, наряду с величинами ASCII соответствующих полей в базе данных. Кроме этого, существует возможность интерактивного доступа к внешнему файлу, словарям кодов и таблице с информацией о процессах, а также к виртуальным управляющим пиктограммам, с помощью которых проблемные элементы могут быть направлены экспертам, помещены в режим ожидания для дальнейшей обработки или переданы в надлежащий массив результатов по переписному участку. Редакторы могут изменять, подтверждать или дополнять данные, разбивать или совмещать изображения страниц и опросных листов, присваивать тот или иной статус записям или направлять результаты по домохозяйствам для обработки конкретными экспертами.

29. Контроль качества редактирования осуществляется после завершения каждой операции редактирования с помощью проверок на полноту, соответствие допустимому интервалу значений и на непротиворечивость. Отбраковка данных в ходе этих проверок ведет к возврату элемента данных или созданию нового элемента для обработки одним и тем же редактором или старшим редактором. Статистические отчеты, созданные в ходе этого процесса, также проверяются контрольными редакторами.

30. В модуле **специального кодирования** осуществляются три более сложные операции кодирования: географическое кодирование, кодирование по занятию и кодирование по отрасли экономической деятельности. Географическое кодирование осуществляется в трех адресных полях: адрес места жительства, адрес места жительства пять лет назад и адрес места работы. Все эти записи подвергаются вводу с клавиатуры и автоматическому кодированию. При невозможности кодирования проблемный элемент передается кодировщику для осуществления кодирования в полуавтоматическом режиме.

31. Подготовка данных к кодированию по признакам "занятие" и "отрасль экономической деятельности" отличается от географического кодирования на стадии ввода с клавиатуры. Поскольку автоматическое кодирование этих полей не предусмотрено процессом кодирования, отсутствует необходимость во вводе текстовых описаний. В процессе полуавтоматического кодирования кодировщик может работать со словами, присутствующими на изображении опросного листа (подробное пояснение см. в разделе "Использование внешних файлов"). Благодаря этому элементы кодирования создаются в ходе

идентификации оптическим считывающим устройством письменного текста в соответствующих полях вопросов; так, например, наличие рукописного текста само по себе является пусковым сигналом для создания элемента кодирования.

32. Интерфейс "кодировщик-машина" имеет удобный для пользователя характер. Кодировщики могут обращаться методом запроса к различным внешним файлам: специальным словарям, файлам работодателей и файлам адресов. Они также могут просматривать изображения отдельных страниц опросного листа и изображения всех страниц, относящихся к соответствующему домохозяйству.

33. Контроль кодирования по отрасли экономической деятельности не является частью системы оптического ввода данных, но определяется ее технологией. Элементы кодирования импортируются из оптического архива, сортируются и посылаются для повторного кодирования на автономные ПЭВМ. Эксперт получает для обработки элемент кодирования в тех случаях, когда первый код, присвоенный кодировщиком ОВД, не согласуется со вторым кодом. Код эксперта заменяет код, созданный ОВД только в том случае, если он отличается от первого кода.

34. По завершении процесса ввода данных создается файл, который отсылается в **оптический архив**. Этот файл содержит все данные с опросных листов, административные данные, статистические отчеты, созданные и использовавшиеся в процессе ввода, а также все данные о проверках каждого поля. Этот файл открыт для доступа и уже используется нами для проведения оценки.

35. Дополнительный **файл**, который отправляется в **центральный компьютер**, содержит полную информацию переписи (без изображений) и данные об обработке каждого поля, которые позволяют определить основные характеристики процесса ввода данных (статус увязки с внешним файлом или статус отмены записи). Эти данные используются для точного и "чувствительного" макроредактирования данных при подготовке окончательных результатов переписи.

Использование внешних файлов

36. Внешние файлы представляют собой файлы различных типов: организационные файлы, регистры, содержащие демографическую информацию, файлы социально-экономических переменных и словари кодов. Они используются для осуществления различных задач по обработке данных, замены компонентов, представления информации и дополнения данных. Внешние файлы могут быть также задействованы для проведения оценки с учетом того, что они не используются в процессе, для оценки которого они должны применяться.

37. Организационные файлы представляют собой перечни различных единиц, наблюдаемых в рамках переписи: перечень всех переписных участков в составе более крупных географических единиц, перечень всех домохозяйств, опрошенных каждым счетчиком

(книга регистрации счетчика), список всего переписного персонала в разбивке по типам функций и т.д. Организационные файлы могут использоваться для решения административных задач, таких, как проверки охвата, проверки полноты передачи материалов, исключительность каждой единицы в системе и мониторинг процесса, мониторинг работы переписного персонала, защита данных и т.д. Они могут также использоваться для повышения качества ввода данных (см. ниже) и разработки статистических отчетов любого вида или рода в системе.

38. Использование регистров, содержащих демографические данные, и файлов социально-экономических данных содействует совершенствованию процесса планирования, а также улучшению охвата и повышению качества ввода данных, редактирования и условных расчетов. Регистр населения Израиля использовался при подготовке всех переписей для целей переписного районирования на основе оценок численности населения. В рамках переписи 1995 года он также использовался для разработки различных слоев географической информации в системе ГИС. Для обеспечения полноты охвата также использовались наклейки, на которых были напечатаны имена и личные идентификационные номера респондентов, по каждому переписному участку. Они оказались весьма полезными, поскольку счетчики с их помощью сверяли, находилось ли соответствующее лицо или нет на ожидаемом переписном участке. Из этого следует, что регистры являются весьма полезными инструментами, даже если не вся содержащаяся в них информация является правильной. В случае Израиля адреса около 70% респондентов были найдены в регистре.

39. Регистр населения также используется в процессе ввода данных. Наличие предварительно напечатанных идентификационных номеров в 70% записей означает, что 70% идентификационных номеров, указанных в опросных листах, будут обрабатываться с нулевым замещением (что означает отсутствие ошибок в оптическом распознавании). Кроме того, увязка между записями регистров и результатами переписи является одной из основополагающих характеристик процесса ввода данных. Она используется для замещения ввода с клавиатуры в следующем процессе: результаты переписи увязываются с соответствующими записями регистра в соответствии с весьма жесткими критериями. Если увязка является успешной, производится сопоставление всех общих переменных. В случае их идентичности величина рассматривается в качестве правильной и это поле полностью пропускается на этапе ввода с клавиатуры. В рамках данного процесса 60% символов, содержащихся в опросном листе (за исключением букв) пропускается при вводе с клавиатуры. Роль регистра в процессе ввода данных является весьма важной, поскольку сокращение длительности этапа ввода с клавиатуры было возможным не только в отношении полей, идентифицированных оптическим считывающим устройством с высокой степенью достоверности, но также и в отношении случаев сомнительного распознавания. В тех случаях, когда величина ОРС подтверждалась величиной регистра, она рассматривалась в качестве правильной без дополнительной обработки.

40. Регистр используется также для ввода данных с тусклых или неразборчиво написанных ответов. Интерактивное обращение к регистру содействовало идентификации величины (но без замены ее величиной регистра). Для этих целей использовался интерфейс "оператор-машина", содержащий изображение опросного листа, соответствующее значение ASCII в базе данных и запись регистра по соответствующему лицу.

41. Этап редактирования в системе ОВД предназначен, кроме прочего, для решения проблем, связанных с невозможностью автоматической увязки записей из файлов переписи и регистра населения. Используя результаты переписи для гибких запросов, редакторы способны идентифицировать искомую запись регистра из всех, предлагаемых для увязки с результатами переписи. Данная увязка записей, производившаяся в полуавтоматическом режиме, позволяла присваивать идентификационные номера записям, в которых отсутствовала данная переменная или содержался неправильный номер. Присвоение проверенного личного идентификационного номера позволяло находить дублирующие записи и исключать их. Кроме того, наличие проверенного идентификационного номера облегчало увязку записи с другими файлами, содержащими одну и ту же переменную идентификационного номера. Благодаря этому процесс редактирования файлов переписи и условных расчетов мог осуществляться на основе внешних файлов, содержащих все или некоторые единицы переписи населения и переменные, подлежащие редактированию или условным расчетам. Внешние файлы также использовались для условного расчета полных записей в качестве источника информации, не запрашивавшейся в опросных листах (вероисповедание и пособия по линии социального страхования).

42. Словари кодов являются основополагающим элементом любого процесса кодирования, независимо от используемой схемы. В компьютеризированной рабочей среде способ автоматического кодирования определяется методом ввода подлежащих кодированию описаний (оптическое распознавание или ввод с клавиатуры).

43. В системах, обеспечивающих оптическое распознавание буквенных символов, первая попытка кодирования может предприниматься в отношении величин, созданных ОРС. Вторая попытка может предприниматься после завершения первого этапа ввода данных, обычно ввода с клавиатуры. В рамках переписи Израиля оптические считывающие устройства не осуществляли распознавания буквенных символов, что означало возможность применения автоматического кодирования только в отношении полей, введенных с помощью клавиатуры. Этот метод использовался в отношении ряда переменных: адрес места жительства пять лет назад, адрес места работы, страна рождения (за исключением семи закрытых категорий) и отношение к основному лицу (за исключением девяти закрытых категорий). Основное правило заключалось в том, что автоматическому кодированию подвергались только недвусмысленные описания переменных, обладающих конечным известным числом категорий. Все другие текстовые описания кодировались с помощью полуавтоматического процесса кодирования (ППК).

44. Основной процедурой процесса ППК является обращение к словарю кодов. Это может производиться на основе использования уже введенных текстовых описаний. Однако изображения опросных листов также являются важными исходными элементами для системы ОВД. Ввод с клавиатуры текстовых описаний подвержен ошибкам и, кроме того, ввод с клавиатуры описаний таких сложных переменных, как занятие или сфера экономической деятельности, является неэффективным в процессе ППК. Более целесообразно производить поиск с помощью ключевых слов, а не всего предложения, содержащегося в соответствующем поле. Изображение опросного листа позволяет произвести это с помощью удобного для пользователя интерфейса. Кодировщики могут гибко производить поиск с помощью слов, которые не были введены в любом другом режиме ввода данных, а не только видеть их в форме изображения. Экономия трудозатрат операторов ввода или кодировщиков, является огромной.

45. Подводя итог, можно сказать, что роль словарей кодов в качестве внешних файлов требует переосмысления в случае их использования в технологической передовой рабочей среде. Что касается кодирования, то в данном случае следует провести оптимизацию всего процесса обработки данных.

46. После проведения переписи населения Израиля 1995 года было принято решение об использовании переписи в качестве источника эмпирических данных для словаря занятий, который будет использоваться для автоматического кодирования. Может создаться впечатление, что решение об отказе от ввода с клавиатуры сведений в процессе регистрации результатов переписи является ошибочным. Однако изображения опросных листов, хранящиеся в оптическом архиве, и величины ASCII соответствующих кодов открывают возможности для оптимизации процесса. Направление запроса в виде цифрового кода в оптический архив позволит получить группу изображений (с использованием специальной технологии), имеющих один и тот же код. Следующей процедурой должен являться ввод с клавиатуры индивидуальных описаний. Это означает, что из всех изображений текстовых описаний, зарегистрированных с помощью одного и того же цифрового кода, вводу с клавиатуры будет подвергаться лишь их небольшое количество. Избежание повторного ввода двух описаний, в которых используется одно и то же слово или комбинация слов, позволяет резко сократить трудозатраты по вводу с клавиатуры всех индивидуальных описаний. В данном случае возможности копирования и переноса изображений служат предпосылками для разработки избирательного процесса ввода данных.

Опросный лист и система ОВД

47. Задаваемые вопросы и их последовательность являются единственными характеристиками опросного листа, которые не зависят от решения об использовании системы ОВД. Все другие характеристики определяются возможностями и ограничениями системы.

Текстура бумаги - способ переплетения волокон бумаги - сжатие, ткацкое соединение или однонаправленное соединение - влияет на впитывание чернил (в виде одной точки или расплывчатого пятна в определенной зоне) и, следовательно, на считываемость записанных величин.

Толщина бумаги - использование слишком тонкой бумаги ведет к возникновению двух основных проблем: захват двух листов при подаче бумаги в сканер и повышенный уровень шума ОРС вследствие прозрачности бумаги. ОРС "видит" символы, написанные на другой стороне листа. Слишком толстая бумага является тяжелой для переноски счетчицами, занимает много места при перевозке в центры ввода данных и может застревать в блоке питания сканирующего устройства.

Плотность чернил - предварительно напечатанные символы должны быть четко видны на одной стороне листа. Если они предназначены для считывания респондентами, они должны быть ясными для них. Если они должны идентифицироваться ОРС, они должны быть машиночитаемыми.

Цвета - в опросном листе используются различные цвета: специальные фоновые цвета, цвета текстовых символов, предназначенных для считывания респондентами, и цвета символов, предназначенных для считывания оптическим устройством.

48. С одной стороны, для компоновки опросного листа более предпочтительно использовать прозрачные цвета (или не использовать их вообще). Исключением служат символы и метки, предназначенные для считывания либо респондентами, либо ОРС. Вычислительные ресурсы всегда являются ограниченными. В случае ввода изображения всей страницы количество регистрируемых пикселей может составить несколько мегабайт, что создаст проблемы с хранением и затруднит передачу информации в рамках системы. С другой стороны, необходимо использовать различные цвета, поскольку вопросы должны быть четкими и понятными, а вопросники выглядеть привлекательными для респондентов, заполняющих их, в особенности если речь идет о методе саморегистрации.

Число страниц - оптическое сканирование только одной стороны листа сопряжено с проблемами, поскольку захват двух листов не может быть автоматически выявлен. При сканировании второй стороны листа может быть активирована процедура контроля (для осуществления контроля сканирования необходимы изображения двух сторон одного листа). Если вопросник состоит из нескольких листов, то каждая сторона каждого листа должна содержать идентифицируемые переменные. Необходимо учитывать, что обработка многостраничных вопросников сопряжено с высокой вероятностью ошибок. Использование таких вопросников ведет к увеличению продолжительности процесса сканирования, а также процесса редактирования, что обусловлено необходимостью идентификации всех структурных единиц файла переписи (опросный лист, домохозяйство и т.д.).

Компоновка страниц - необходимо отдельно производить сканирование каждой индивидуальной страницы. Это означает, что вопросник должен состоять либо из отдельных страниц, либо из сшитых страниц, которые должны разъединяться перед сканированием. В первом случае возникают проблемы, связанные с потерей страниц, передачей страниц из разных опросных листов одному домохозяйству, а также с использованием страниц одного и того же вопросника для опроса различных домохозяйств. Во втором случае требуется дополнительный персонал для расшивки и компоновки вопросников перед сканированием. Страницы должны быть расшиты заблаговременно во избежание попадания пыли в сканер, поскольку ее попадание в считывающее устройство может привести к техническим неполадкам и затрудняет процесс оптического распознавания. Поля должны быть достаточно широкими, для того чтобы расшивка страниц не сказалась отрицательно на письменных ответах.

Элементы графики - существует несколько причин для включения графических элементов в опросные листы: облегчение понимания опросных листов респондентами, повышение эффективности процесса сканирования и оптического распознавания символов (ОРС).

49. Респонденты должны четко определять начало и конец вопроса, последовательность вопросов, поля для заполнения, а также поля, предназначенные для представления информации по каждому члену домохозяйства.

50. Что касается облегчения процесса сканирования, то линии или рамки служат упорядочению компоновки страниц, что имеет чрезвычайно важное значение для открытия окон оптического распознавания в точном соответствии с изображением опросного листа. Также необходимо обеспечить точную ориентацию каждой страницы в соответствии с шаблоном. В данном случае даже частичное сохранение предварительно напечатанных вопросов и пометок ведет к нерациональному использованию вычислительных ресурсов.

51. Поскольку система ОРС работает с предварительно определенными полями, необходимо также четко определить границы окна распознавания, открываемого в ходе процесса. Размеры данного окна должны быть гибкими (поле плюс согласованные пробелы), поскольку в некоторых случаях респонденты стирают ответы или пишут их сверху или ниже. Необходимо также обеспечить обособление соседних полей. Это означает, что пробелы между полями опросного листа должны быть достаточно широкими для обеспечения гибкого ОРС, но не слишком широкими для того, чтобы это привело к увеличению числа страниц вопросника.

Вопросы, допускающие единственный ответ - оптическому считывающему устройству намного проще идентифицировать пометку, чем символ. Намного легче и надежнее определить что-то написанное, чем то, что точно написано. Если планируется использовать оптическую систему, то более целесообразно включить в опросный лист как можно большее число вопросов, допускающих единственный ответ, однако без нанесения ущерба их полноте и надежности.

Использование предварительно напечатанных полей - если использование меток невозможно, то следует учитывать, что ОРС лучше справляется с предварительно определенными формами символов. Предварительно напечатанные с использованием определенного вида и размера шрифта символы менее подвержены ошибкам при считывании. С учетом этого целесообразно напечатать в опросном листе все заранее известные переменные. Примером таких переменных могут служить номер вопросника, номер страницы и номер записи. В некоторых случаях предпочтение отдается печатанию этих номеров на наклейках ввиду непредсказуемых факторов. Так, например, цифра, используемая для комбинирования двух вопросников по одному домохозяйству, должна печататься на наклейке, поскольку решение об ее использовании будет приниматься в ходе процесса регистрации, а не до него. Номер переписного участка может быть предварительно напечатан на вопроснике, однако это ведет к нерациональному использованию опросных листов, поскольку не все они могут быть использованы на определенном переписном участке. Наклейка с личным идентификационным номером будет прикрепляться к вопроснику в том случае, если соответствующее лицо будет находиться по предполагаемому адресу и т.д.

Добавление "системных" переменных - в условиях компьютеризированной рабочей среды возрастает потребность в идентификации каждой физической и логической единицы. В контексте системы ОВД были добавлены переменные для идентификации одного листа (для правильного исключения и ОРС), двух сторон одного и того же листа, всех страниц одного и того же вопросника, всех вопросников по одному и тому же домохозяйству, всех домохозяйств, входящих в состав одного переписного участка, и всех составляющих каждой индивидуальной записи.

Инструкции по заполнению - с учетом растущего нежелания респондентов сотрудничать с переписным персоналом опросные листы должны быть ясными для понимания и содержать немногочисленные, но простые инструкции по их заполнению. Адаптация опросного листа к требованиям системы оптического считывания является задачей переписных органов, а не респондентов. Если перепись проводится методом опроса респондентов счетчиками, инструкции могут быть расширены. Решение об использовании прописных или строчных букв для заполнения опросного листа зависит от уровня сотрудничества между счетчиками и респондентами. Использование прописных букв облегчает процесс оптического распознавания и содействует повышению надежности информации, однако требует больше времени и самодисциплины. Следовательно, если можно надеяться на получение ответов, написанных прописными буквами, будет целесообразно включить в опросный лист такую инструкцию. Если просьба о заполнении опросного листа прописными буквами может негативно сказаться на сотрудничестве с респондентами, следует избрать другую стратегию.

52. Благожелательное отношение, дисциплинированность респондентов и точность заполнения ими опросных листов зависят от их культуры, в связи с чем соответствующее решение должно приниматься с учетом местных условий, а не носить глобальный характер.

В случае Израиля получение ответов, написанных прописными буквами, представляется маловероятным, однако в иврите буквы пишутся раздельно, в связи с чем они легче поддаются идентификации по сравнению с английским языком. В рамках переписи 1995 года буквенные символы не считывались ОРС, поскольку анализ выгод и затрат дал неудовлетворительные результаты. В то же время производилось оптическое распознавание цифровых символов.

Заключительные замечания

53. Наш вывод заключается в том, что оптимальным способом использования новых технологий для целей переписи является переосмысление и изменение формулировки процессов переписи и промежуточных целей. Новые технологии уже доказали и еще, вероятно, неоднократно докажут свою перспективность и полезность с точки зрения совершенствования процессов переписи, рационального использования ресурсов и повышения качества материалов переписи.

54. Однако ни одна из технологий не может быть автоматически применена без проведения стандартных проверок качества и корректировок с целью обеспечения требуемого уровня качества. Непрерывный и сквозной контроль качества на протяжении всего процесса является одним из обязательных условий, которое относительно легко выполнить в рабочей среде изображений в формате ASCII.

55. Компьютеризация процессов позволяет значительно сократить масштабы участия в них людей, однако машина еще не может полностью заменить человека. Некоторые задачи по проведению переписи будут осуществляться в системе взаимодействия "человек-машина", обычно в виде полуавтоматических процедур.

56. В будущем уже, вероятно, заложены зачатки новых технологических разработок. Однако нам еще предстоит внести технологические усовершенствования в планирование переписи с учетом уже существующих технологий. Наиболее очевидной задачей является интеграция других процедур переписи, начиная со сбора данных и кончая разработкой материалов переписи, в эти передовые технологические системы. Система ОВД, которая была разработана для переписи населения Израиля 1995 года, была призвана обеспечить интеграцию процедур ввода данных. Однако ее характеристики оказали положительное влияние также и на операции по сбору сведений, осуществляемые до ввода данных, и на операции обработки и анализа данных, выполняемые после него. Систематическая и тщательная разработка интегрированной системы позволит дополнительно расширить возможности использования потенциала уже существующих технологических средств.

57. Дополнительная информация о системе ОВД, использовавшейся при проведении переписи населения и жилищ Израиля 1995 года, **может быть найдена в следующих публикациях**:

Blum, Olivia, "Editing Definition and Operation in the Optical Data Entry System (ODE) of the 1995 Census of Population in Israel". Working paper No. 48. SCECE Work Session on Statistical Data Editing. Athens. Nov. 1995.

Blum, Olivia, "Evaluation of Data-Editing Using Administrative Records". Working paper No. 16 SCECE Work Session on Statistical Data Editing. Prague. Oct. 1997.

Blum Olivia & Ben-Moshe, "Automated Record Linkage and Editing: Essential Supporting Components in Data Capture Process". In Statistical Policy Working Paper No. 25. Data Editing Workshop and Exposition. Executive Office of the President of the United States. Washington DC. December 1996.

Nathan Gad & Israel Givol, "The ODE (Optical Data Entry) Experience in Israel". Paper presented at InterCasic'96', San Antonio, Texas. December 1996.

Следующие документы были опубликованы в серии "Новые технологии для проведения цикла переписей 2000 года", Европейско-Средиземноморское рабочее совещание в Израиле, март 1997 года:

Blum, Olivia, "The ODE System: Logical Structure and Guiding Principles".
Blum, Olivia, "Keying Module".
Blum, Olivia, "Editing and Coding Module".
Givol, Israel, "The ODE System: Technology and Human Engineering.
Kagan, Oren, "The World of Scanning".
Khomenko, Yuri, "ODE Telecomputing Systems".
Yitzhaki, Ruhama, "The Operational System of the ODE".
