Технология ввода данных с бумажных документов

М.М.Мучник, к.т.н., с.н.с., доцент

МСУ, компания “Сканпрог”, 04070, Украина, Киев, вул.Фроловська, 6/8, корп.1, Факс: (380-44) 463-76-14,
Тел.: (380-44) 463-76-15, E-mail: muchnik@padco.kiev.ua

Аннотация. В докладе рассмотрены технологии ввода данных с бумажных носителей и предложена альтернативная технология ввода данных и программный комплекс поддержки ее этапов Imagewriter , проведен сравнительный анализ этих технологий.

Вступление

В течениии многих столетий основным носителем информации, используемым как для подготовки различных документов, так и для их обработки и хранения, была бумага. И в наше время, несмотря на все расширяющееся использование компьютеров практически во всех сферах человеческой деятельности, бумага остается одним из основных носителей, используемых для подготовки и хранения документов. Все большее количество документов сразу готовится на электронных носителях с помощью компьютера, однако использование бумажных носителей информации не утрачивает своего значения, и, в первую очередь, в связи с простотой их **подготовки.** Для работы человека с бумажными документами достаточно нескольких листов бумаги и ручка - и больше никаких дополнительных устройств. Однако для **обработки** данных обычно используется компьютер, поскольку эффективность обработки на бумажных носителях очень низка.

Во многих информационных технологиях в настоящее время бумажные носители используются для подготовки исходных документов, визуализации и придания легитимности отчетам (подписями и печатями), а для обработки и хранения данных - компьютеры, т.е. постоянно возникает необходимость ввода данных в компьютер с бумажных носителей (формирование электронных версий данных). Это довольно трудоёмкий процесс, от производительности которого зависит производительность информационной технологии в целом.

В настоящей работе рассмотрены известные технологии ввода данных с бумажных носителей, описана альтернативная технология ввода данных и программный комплекс поддержки ее этапов **Imagewriter,** проведен сравнительный анализ технологии **Imagewriter** с уже известными ранее технологиями ввода данных.

Для дальнейшего изложения введем несколько определений:

- Поле на бумажной форме или ее графическом образе, содержащее данные для ввода, будем называть **«Полем считывания данных»,** а поле на экране компьютера и/или в памяти компьютера, в которое вводятся данные–**«Полем ввода данных»**;
- Если данные вносятся в бумажные формы вручную (вписываются от руки карандашом или ручкой), будем говорить о **ручных** методах и средствах подготовки данных;
- Если данные вносятся в бумажные формы с помощью механического или електро-механического устройства (например, пишущей машинки), будем говорить о **механизированных** методах и средствах подготовки данных;
- Если подготовка и/или ввод в компьютер выполняются человеком-оператором для каждого отдельного элемента данных (т.е. с существенной долей ручного труда), будем говорить об **автоматизированных** методах и средствах.;
- Если набор (массив) данных подготавливается с помощью компьютера и/или вводится в компьютер с минимальным участием человека, состоящим в инициализации программы подготовки или ввода данных (или без его участия), будем говорить об **автоматических** методах и средствах. Например, автоматическим является ввод массива данных в память компьютера из заранее подготовленного файла.

Типовые этапы технологического процесса ввода данных в компьютер с бумажных форм (документов), выполняемые независимо от типа используемых методов и средств ввода данных, приведены в столбце 2 таблицы 1. Для сокращения текста в других столбцах этой таблицы используется набор символов <...>, обозначающий вставку текста из столбца 2 данной строки таблицы.

Таблица 1. Этапы технологического процесса ввода данных в различных технологиях

№ п/п	**Типовые этапы ввода данных**	**Автоматизированный ввод с бумажных форм**	**Автоматическое сканирование, автоматизированный ввод с образов бумажных форм (Imagewriter)**	**Автоматическое сканирование и автоматические распознавание и ввод в память**
1	2	3	4	5
1	Сортировка бумажных документов и разделение их на пачки	Ручная <...>	Ручная <...>. Автоматическое или автоматизированное сканирование пачек документов (формирование каталогов файлов образов документов)	Ручная <...>. Автоматическое сканирование пачек документов (формирование каталогов файлов образов документов)

2	Выбор и подготовка пачки бумажных документов для ввода	Ручной <...>	Автоматизированный выбор и открытие каталога образов документов для ввода	Автоматизированный выбор и открытие каталога образов для ввода
3	Выбор из пачки очередного документа для ввода	Ручной <...>	Автоматический или автоматизированный выбор из каталога очередного образа для визуального считывания данных	Автоматический выбор из каталога очередного образа для автоматического распознавания.
4	Определение типа выбранного документа и подготовка программы к вводу данных этого типа	Визуальное <...>	Автоматизированное определение типа очередной формы и наложение шаблона на ее образ	Автоматическое распознавание типа очередной формы и наложение шаблона на ее образ
5	Поиск очередного поля считывания данных	Визуальный<...>на бумажной форме	Автоматический <...> на образе формы и его активизация	Автоматический <...> на образе бумажной формы
6	Считывание и запоминание текста, содержащегося в текущем поле считывания данных	Визуальное <...> бумажной формы	Визуальное <...> на образе бумажной формы	Автоматическое <...> на образе бумажной формы
7	Распознавание считанных данных и их ввод в текущее поле ввода данных	Визуальное <...> с помощью клавиатуры компъютера	Визуальное <...> с помощью клавиатуры компъютера	Автоматическое <...> в памяти (автоматически). Формирование списка неуверенно распознанных символов
8	Сравнение введенного текста с исходным. Если есть несоответствия, то его редактирование	Визуальное <...>	Визуальное <...>	
9	Автоматическая проверка корректности введенных в поле ввода данных.	<...> Если данные некорректны, то их редактирование	<...> Если данные некорректны, то их редактирование	<...> Формирование списка некорректных данных
10	Проверка окончания ввода данных с текущего документа. Если текущее поле – не последнее на странице, то переход к п.5	<...>	<...>	<...>
11	Автоматическая проверка взаимной корректности данных между полями текущего документа. Если данные некорректны, то их редактирование	<...>	<...>	Автоматизированная верификация неуверенно распознанных и некорректных данных. <...>
12	Автоматический экспорт введенных по текущему документу данных в файл или базу данных	<...>	<...>.	<...>
13	Проверка окончания пачки документов. Если текущий документ – не последний в пачке, то переход к п.3	Ручная <...>	Автоматическая или автоматизированная <...>	Автоматическая <...>
14	Проверка наличия рассортированных пачек документов. Если текущая пачка – не последняя, то переход к п.2	<...>	<...>	<...>
15	Проверка наличия нерассортированных по пачкам документов. Если есть, то переход к п.1	<...>	<...>	<...>
16	Передача пачек введенных документов в архив и окончание ввода данных	<...>	<...>	<...>

1. Технология автоматизированного ввода

Информационная технология автоматизированного ввода предполагает ввод оператором данных, записанных на бумажных формах, посредством клавиатуры компъютера и с помощью специально подготовленной программы ввода. Каждому полю считывания данных на бумажной форме соответствует поле ввода данных на экране и в памяти компъютера, а бумажная и экранная формы эквивалентны не по виду, а по наборам полей. Компъютер с инсталлированной программой ввода называют автоматизированным рабочим местом (АРМ) ввода данных [1], хотя обычно АРМ ввода данных входит в состав АРМ‘а более общего назначения, выполняющено наряду с вводом и ряд других функций, например, обработку и формирование отчетов. Этапы технологического процесса автоматизированного ввода представлены в столбце 3 таблицы 1.

Скорость автоматизированного ввода данных довольно низка как по причине низкой скорости печатания оператором на клавиатуре, так и в связи с необходимостью значительных затрат времени на вспомогательные действия по обеспечения ввода с бумажных документов. Опытные операторы обычно работают двумя руками «слепым десятипальцевым методом» [2], т.е. смотрят только на бумажный лист, не переводя взгляд на клавиатуру и/или экран компъютера для контроля. В результате появляется достаточно много ошибок в данных. Если в программе ввода предусмотрены правила контроля введенных данных, то при вводе неверного символа будет выдано соответствующее диагностическое сообщение и оператор будет вынужден приостановить процесс ввода и провести редактирование введенного текста, хотя это замедляет процесс ввода. Если такая корректировка не производится, то данные передаются на следующий этап обработки с ошибками и их исправление требует в дальнейшем достаточно больших трудозатрат.

Несмотря на все указанные недостатки автоматизированные рабочие места (АРМ) ввода данных, представляющие собой персональный компъютер с программой автоматизированного ввода, в настоящее время широко используются для ввода данных с любых бумажных форм, заполненных визуально распознаваемым текстом. Разработка программного обеспечения АРМ‘а ввода относительна проста и не требует больших временных, финансовых и трудовых затрат (для одной формы АРМ может быть разработан за 0.5-2 месяца).

Требования к качеству бумажных форм, используемых при технологии ручного ввода, являются достаточно низкими. Обычно такие формы печатаются по принципу наименьших затрат, т.е. на газетной бумаге, черной краской и на типографском оборудовании ближайшей типографии (обычно, довольно старом). Если формы закончились, то их можно напечатать на любой типографской или ксерокопировальной технике.

Таким образом, создание средств автоматизированного ввода не требует больших затрат, эта технология проста в эксплуатации, однако она не позволяет обеспечить высокую производительность (опыт показал, что за одну смену оператор автоматизированно вводит данные не более чем с 60-100 форм средней заполненности).

2. Технология сканирования и автоматического распознавания

Информационные технологии сканирования и автоматического распознавания текстов (OCR для печатных и ICR для печатных и рукописных текстов) предполагают предварительное сканирование бумажных форм с последующим автоматическим распознаванием графического образа текста всего документа либо отдельных его полей (полей считывания), выделенных в соответствии с заранее подготовленным шаблоном и запись распознанных данных в поля ввода в памяти компъютера. Современные OCR обеспечивают распознавание большинства типов печатных документов, заполненных сплошным текстом и поэтому могут использоваться достаточно широко. Однако главной проблемой остается ввод данных с бумажных форм, содержащих поля, заполненные рукописными текстами. Такие формы, называемые машиночитамыми, разрабатываются специально с учетом требований сканирования и автоматического распознавания с помощью ICR, печатаются в одной типографии на качественной бумаге и аккуратно заполняются (например, формы для переписей населения). Для подготовки графических образов документов в основном используются сканеры с автоматической подачей бумаги с лотка (автоматическое сканирование). В столбце 5 таблицы 1 приведены этапы технологического процесса сканирования и автоматического распознавания. Как видно из этого перечня, почти все этапы, кроме подготовки документов для сканирования и верификации, выполняются автоматически, т.е. не для одного, а для группы документов.

Современные промышленные сканеры позволяют сканировать до 150-200 стр/мин, современные компъютеры обеспечивают распознавание 40-60 форм/мин, однако один верификатор может обработать от 300 до 500 форм за смену, т.е. около 1 формы в минуту. Поскольку в современных ICR обычно обеспечивается возможность одновременной работы нескольких станций верификации, использование ICR действительно позволяет существенно ускорить формирование электронных версий бумажных документов. В то же время, учитывая то, что этап верификации выполняется автоматизированно, более правильным будет говорить не о автоматической работе комплекса ICR, а о том, что по сравнению с программами автоматизированного ввода в комплексах ICR большее число этапов выполняется автоматически.

Тем не менее, говорить о повсеместной замене автоматизированного ввода сканированием и автоматическим распознаванием пока не приходится, поскольку технология ICR эффективна только для распознавания данных, подготовленных на машиночитаемых формах (лучше всего- заполненных числовыми данными). При автоматическом распознавании текстовых данных достоверность распознавания обычно снижается, что замедляет процесс верификации до 200-250 форм за смену.

Для обеспечения более высокого качества распознавания, машиночитаемые формы должны быть не черно-белыми, а цветными, обычно красно-оранжевыми (этот цвет хорошо отсекается при сканировании) и

совпадать на просвет. Цветные формы дороже обычных черно-белых и их нельзя ксерокопировать на обычной ксерокопировальной технике, когда пустые формы заканчиваются. В отдаленных от центра районах это часто становится серъёзной проблемой.

Стоимость как отдельных компонентов, так и всего комплекса ICR в целом, довольно высока. Например, стоимость лицензии на использование одного АРМ‘а системы Eyes&Hands FORMS шведской фирмы ReadSoft составляет 5200 евро, а системы FormReader российской компании ABBYY – $1200 за распознавание 10000 страниц в месяц. Поэтому комплексы ICR обычно размещают в специальных центрах сканирования, рассчитанных на обработку десятков тысяч документов в день. В свою очередь, это требует организации специальной службы доставки документов и организации их последущего возврата и/или архивации.

Итак, для использования технологий автоматического распознавания требуется предварительно выполнить следующие подготовительные работы:

- разработать машиночитаемые формы, регулярно печатать их в достачном количестве в одной типографии и рассылать для заполнения юридическим или физическим лицам, которые должны готовить и сдавать отчеты по этим формам;
- закупить оптический сканер (один или несколько), компъютеры, сетевое оборудование и программное обеспечение для создания комплекса (ов) сканирования и автоматического распознавания;
- разработать шаблоны машиночитаемых форм и настроить комплекс ICR на их распознавание;
- разработать программное обеспечение для интеграции результатов распознавания в конкретную информационную технологию;
- обеспечить обучение операторов и вспомогательного персонала для поддержки комплекса в работоспособном состоянии;
- обеспечить регулярное финансирование работ по эксплуатации комплекса.

Таким образом, использование технологий ICR обеспечивает ускорение ввода данных со специально разработанных и акуратно заполненных бумажных форм, но требует достаточно больших финансовых затрат как при подготовке и первичном внедрении этих технологий, так и в процессе их эксплуатации. При распознавания черно-белых форм, заполненных рукописным текстом, эффективность ICR резко снижается.

3. Технология сканирования и автоматизированного распознавания (Imagewriter)

В декларативном патенте [3] был предложен способ ввода информации в память компъютера с графического образа бумажного документа, при котором в одном окне на экране оператор видит выделенный отличным от общего фона формы цветом фрагмент графического образа бумажного документа, содержащий данные для считывания (поле считывания данных), а в расположенное в непосредственной близости от первого (обычно, под ним) другое окно (поле ввода данных), которое пусто, с помощью клавиатуры (автоматизированно) вводит данные, визуально считанные им из первого окна. После окончания ввода в окно ввода данных оно закрывается, введенные данные записываются в память, а на экран выводится образ следующего поля считывания данных и пустое поле ввода данных и т.д. Координаты полей считывания данных на образе документа и их последовательность задаются (выбираются) в соответствии с заранее созданным графическим шаблоном. В каждом поле ввода есть заголовок с его названием. На рисунке 1 представлен фрагмент образа бумажной формы, на котором выделены поля считывания и ввода данных. Графические образы бумажных документов формируются путем их сканирования .

В столбце 4 таблицы 1 представлены этапы технологии автоматизированного формирования электронных версий бумажных документов **Imagewriter,** разработанной на основании описанного способа и реализованной в одноименном программном комплексе, состоящем из трех основных взаимосвязанных компонентов: редактора шаблонов, редактора словарей и собственно программы ввода.

Редактор шаблонов позволяет описать для каждого типа бумажной формы графический шаблон (с помощью специального встроенного графического редактора), типы полей считывания и ввода (см. Табл.2), а также правила контроля корректности введенных данных (описываются на встроенном языке описания правил контроля корректности данных). .

Таблица 2. Типы полей считывания и ввода

Тип поля считывания	**Тип поля ввода**	**Тип поля в базе данных**
Текстовое	Текстовое поле, расширенное текстовое поле, текстовое поле с маской, простой список, комбинированный список Поле для ввода даты	Символьный Дата
Отметка	Флажок	Логический
Группа отметок	Радиокнопки, простой список, комбинированный список	Символьный
Графический объект	Картинка	Символьный

Если в некотором поле считывания могут быть записаны только стандартные значения из ограниченного набора, то список этих значений может быть заранее оформлен в виде словаря, который при описании шаблона связывается с этим полем. Формирование словарей реализуется с помощью **Редактора словарей,** позволящего для каждого значения в поле считывания указать значение в поле ввода (которое и будет экспортировано). Эти

значения могут совпадать, но могут и отличаться. Например, если в поле считывания указываются имена месяцев, то в базу данных лучше экспортировать номер месяца – он занимает меньше места и с ним проще работать.При вводе данных оператору достаточно выбрать из списка необходимое значение, которые и будет записано в поле ввода. Использование словарей в ряде случаев существенно ускоряет получение значения в поле ввода без поэлементного ввода текста с клавиатуры.

Программа ввода (собственно Imagewriter) выполненяет следующие функции:

- указание параметров для первичной настройки комплекса;
- управление сканированием бумажных документов и формирование каталога их образов;
- открытие каталога ранее отсканированных документов;
- автоматизированный (автоматический) выбор из каталога образа документа для ввода;
- выбор вида представления образа документа на экране;
- увеличение или уменьшение масштаба образа документа на экране;
- поворот образа документа;
- автоматизированный (автоматический) выбор и наложение шаблона на образ документа;
- автоматизированный ввод данных в поля ввода на образе документа;
- автоматический контроль корректности введенных данных по правилам контроля;
- автоматический экспорт введенных данных.

В качестве параметров, в частности, указываются формат экспорта введенных данных (в настоящее время - CSV, DBF и XML) и язык интерфейса пользователя (в настоящее время - украинский, русский и английский). В настройке на национальный язык ввода данных Imagewriter не нуждается, поскольку в нём, как и в любой другой прикладной программе Windows, используется стандартная программа ввода с клавиатуры.

После окончания ввода данных в поле ввода цвет соответствующего поля считывания в шаблоне изменяется на салатовый и в дальнейшем сохраняется. Это существенно облегчает визуальный контроль результатов ввода данных по образу документа, поскольку при открытии документа сразу видно, в какие поля данные вводились, а в какие – нет. Такая возможность особо интересна для контроля документов, заполненных в основном полями считывания типа «отметка», поскольку цвет определяет и значение такого поля, что позволяет очень просто визуально проверять как полноту, так и правильность ввода данных оператором.

Визуальный контроль значений текстовых полей может проводиться либо в процессе ввода данных либо после ввода данных путем просмотра документов и активации полей считывания, поскольку введенные значения сохраняются в каталоге графических образов документов.

В Imagewriter‘е предусмотрен специальный режим просмотра введенных данных, в котором эти данные не исчезают при переходе к следующему полю, а остаются видимыми на экране в полях считывания. Это позволяет оператору верифицировать содержимое всех полей документа без предварительной активации. Imagewriter также позволяет распечатать форму с заполненными введенными данными полями, либо впечатать введенные данные в пустой бланк нужной формы.

Следует отметить, что средства переноса данных из одного поля на экране компьютера в другое поле использовались в некоторых системах и ранее, однако они всегда рассматривалось как дополнительные и их не пытались сделать более эффективными для обеспечения массового ввода. В комплексе Imagewriter, по мнению автора, эти средства реализованы в достаточно удобном для ввода и контроля данных виде.

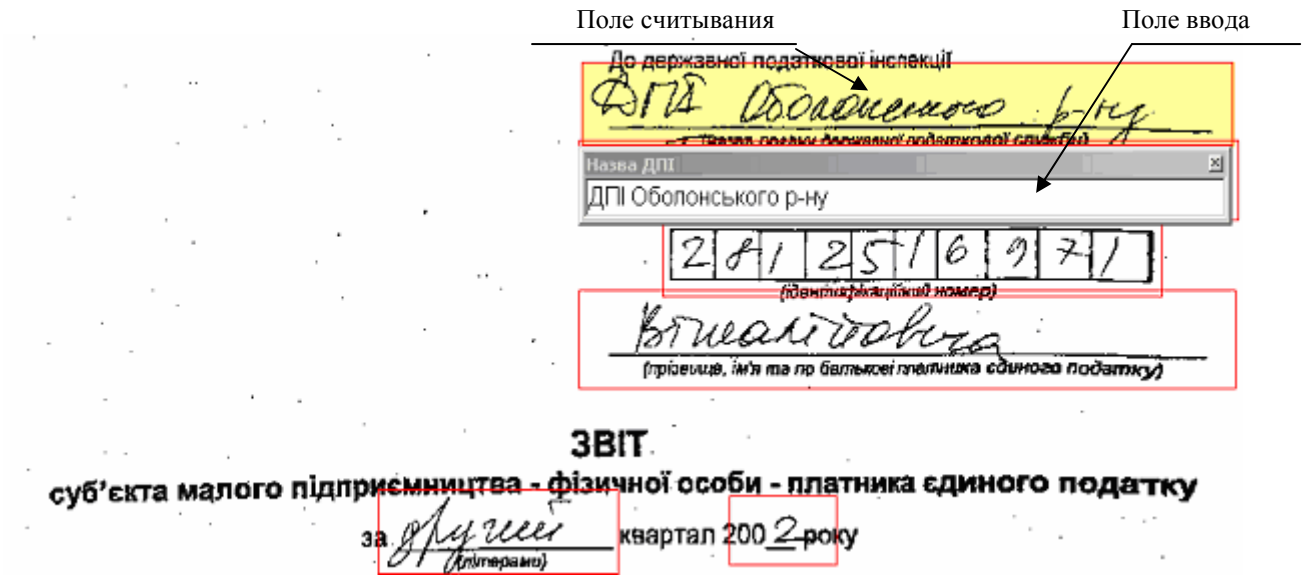

Рисунок 1. Поля считывания и ввода на графическом образе бумажной формы

4. Сравнительный анализ технологий ввода данных

Проведем сравнительный анализ различных технологий ввода данных, представленных в таблице 2. Для удобства ссылок обычную технологию автоматизированного ввода с бумажных форм будем называть «технология автоматизированного ввода», технологию сканирования и автоматизированного ввода с образа бумажной формы на экране компъютера - «технология Imagewriter", а технологию сканирования и автоматического распознавания – «технология автоматического распознавания».

Сравнение «технологии автоматизированного ввода» и «технологии Imagewriter». Общим для этих технологий является то, что распознавание содержимого полей считывания оператор выполняет визуально, а ввод данных в поля ввода на екране компъютера производит с помощью клавиатуры.

К различиям следует отнести то, что в технологии Imagewriter производится предварительное сканирование как пустых бумажных форм документов (для создания шаблонов), так и заполненных бумажных форм документов (для собственно ввода данных). Благодаря этому обеспечивается возможность повышения производительности труда оператора в процессе ввода данных за счет исключения необходимости оперирования с бумажными документами (перенос со склада к рабочему месту пачки документов и ее передача в архив, поиск и перемещение каждого документа после окончания ввода и др.), а также появляется новая возможность: обработка графических образов (чертежей, рисунков, фотографий). Для полей типа «графический образ» Imagewriter экспортирует ссылку на файл, в котором содержится этот образ. Это позволяет использовать технологию Imagewriter для создания архивов чертежей, картин или любых объектов, содержащих либо только графический образ, либо графический образ и текстовую информацию. Кроме того, поля считывания и ввода в обычной технологии автоматизированного ввода располагаются в разных местах (на бумажной форме и экране компъютера соответственно), что вызывает необходимость затрат времени на визуальный поиск этих полей оператором, а в технологии Imagewriter – оба поля находятся рядом на экране компъютера, а их поиск и фокусировка на них взгляда оператора обеспечиваются самой системой (поле считывания выделено желтым цветом), что существенно ускоряет процесс ввода.

Важным достоинством технологии Imagewriter является также отмеченная выше возможность быстрого визуального контроля полноты ввода данных по цвету полей шаблона, а для полей типа «отметка» - и правильности введенных значений.

Для проведения сравнительного анализа стоимости закупки программмного обеспечения (ПО) для обоих технологий, необходимо знать стоимость ПО АРМ автоматизированного ввода. Однако заранее определить стоимость АРМ автоматизированного ввода для всех случаев практически невозможно, поскольку такие АРМ‘ы всегда являются специализированными, а не универсальными программами, разрабатываются по заказу для ввода данных четко определенного типа и в них обычно, как отмечалось выше, входят не только программы ввода, но и обработки данных, а также средства формирования отчетов. В связи с этим в настоящей работе стоимость закупки ПО для обоих технологий будем считать эквивалентной и не влияющей на стоимость внедрения и эксплуатации. Для анализа экономической целесообразности использования технологии Imagewriter в конкретных приложениях, такой расчет должен быть проведен обязательно.

Что касается сравнительной стоимости подготовки, внедрения и эксплуатации апаратных средств этих технологий, то в них обоих в качестве рабочих мест операторов используются персональные компъютеры, объединенные локальной сетью. В технологии Imagewriter необходим также и оптический сканер. В настоящее время во многих организациях уже есть оптические сканеры или документ-центры, в состав которых входит оптический сканер. При отсутствии последнего планшетный сканер можно приобрести за $50-150, а офисный сканер с автоматической подачей бумаги формата А4 – за $500-1500 в зависимости от производительности (10-30 стр./мин) и фирмы-изготовителя, что вполне сравнимо со стоимостью одного компъютера.

Производительность труда оператора при использовании технологии Imagewriter возрастает в 1.5-2 раза, что означает, что закупка сканера для использования технологии Imagewriter становится экономически выгодной, если ввод данных выполняет более 2-х операторов, поскольку затраты на эксплуатацию обоих технологий практически эквивалентны (это, в основном, зарплата операторов).

Особо следует отметить возможность использования программного комплекса Imagewriter для разработки АРМ‘ов автоматизированного ввода данных непосредственно с заполненных бумажных документов (без их предварительного сканирования). Вообще говоря, генерация таких АРМ‘ов планировалась как отдельная фунция Imagewriter (реализация этой функции будет выполнена до конца 2004г.), однако в процессе разработки стало ясно, что уже сейчас Imagewriter обеспечивает создание таких АРМ‘ов более простым путем:

- вначале разработчик шаблона сканирует одну пустую бумажная форму и, с помощью редактора шаблонов, формирует её графический шаблон;
- перед началом ввода оператор создаёт каталог файлов графических образов пустых форм путем копирования в необходимом количестве графического файла образа формы, использовавшегося для создания шаблона (без сканирования);
- сформированный шаблон накладывается на образы пустых форм в каталоге;
- из очереди образов пустых форм с наложенными шаблонами оператор последовательно выбирает очередной образ и производит ввод данных в поля ввода на экране таким же образом, как и в обычном режиме работы Imagewriter, только считывание данных оператор производит непосредственно с бумажных форм, а не с образа на экране;
- по окончании ввода данные экспортируются.

Конечно, скорость ввода данных в описанном режиме практически не будет отличаться от скорости обычного автоматизированного ввода, однако на экране оператор будет видеть образ той же самой формы, что и на бумаге (это более наглядно), а разработка такого АРМ‘а с помощью редактора шаблонов занимает очень короткое время (1-3 часа в зависимости от количества полей формы и квалификации разработчика шаблона).

Таким образом, технология Imagewriter обладает всеми достоинствами технологии автоматизированного ввода (относительно низкая стоимость внедрения и эксплуатации, возможность работы с любыми бумажными формами) и позволяет резко сократить время и трудозатраты на разработку АРМ‘ов ввода данных, сделать их нагляднее для операторов, а также ускорить процесс ввода данных за счет автоматизации вспомогательных операций в 1.5-2 раза при использовании сканера для создания образов заполненных форм.

Сравнение «технологии автоматического распознавания» и «технологии Imagewriter». Общим для этих технологий является то, что пачки бумажных форм предварительно сканируются и выделение полей считывания производится в соответствии с заранее разработанным шаблоном. Отличия состоят в методах распознавания и ввода данных в поля ввода: в технологии Imagewriter – визуальное распознавание и автоматизированный ввод, в технологии автоматического распознавания – автоматическое распознавание и автоматический ввод, а также в подходе к верификации данных.

Возможность автоматического распознавание и ввода является несомненным достоинством. Для машиночитаемых документов это действительно позволяет получить значительную экономию времени и трудозатрат. Однако для обычных черно-белых форм, заполненных слитным текстом, технология автоматического распознавания практически не применима. Для таких форм достоинством становится возможность визуального распознавания, использующуяся в технологии Imagewriter и применимая для любых форм и текстов. Именно такие формы в основном используются в документообороте большинства организаций.

Следует отметить, что структура программного комплекса Imagewriter такова, что позволяет подключить к нему модуль автоматического распознавания печатных и/или рукописных символов. В этом случае его можно будет рассматривать как достаточно развитый комплекс ICR. Но тогда потребуется выдвинуть соответствующие жесткие требования к бумажным формам, да и стоимость комплекса резко возрастет, т.е. будут потеряны те преимущества, которые дает визуальное распознавание. В то же время существует достаточно много документов, в которых есть поля, хорошо распознаваемые автоматически (например, содержащие напечатанный текст) и поля, автоматически не распознаваемые. Для обработки таких документов необходимо не автоматическое распознавание всех полей, а возможность автоматического распознавания поля по указанию оператора (например, нажатие комбинации функциональных клавиш). В этом случае оператор сам визуально определит, насколько возможно автоматическое распознавание текста из поля считывания и либо вызовет программу автоматического распознавания либо сам введет этот текст с клавиатуры. Такой подход позволит обеспечить быструю обработку значительно большего числа типов документов, чем это позволяют известные системы автоматического распознавания. Пользователь получит возможность выбирать между менее дорогой и менее производительной или более дорогой и более производительной системами.

Верификацию в технологии Imagewriter можно проводить визуально как сразу после ввода данных в очередное поле ввода текущей формы, так и по окончанию обработки страницы или группы страниц. В технологиях автоматического распознавания обычно проводится групповая верификация в виде отдельного этапа (этапа верификации), т.е. по всем полям текущей формы либо по всем полям всех распознанных документов, что позволяет специализировать операторов именно для этого вида работ. Кроме того, в комплексах ICR при распознавании обычно отмечаются неуверенно распознанные символы и на верификацию оператору выдаётся на экран не весь распознанный текст, а только неуверенно распознанные символы. Преимущества технологии автоматического распознавания в этом очевидны. Однако в тех случаях, когда некоторый символ распознан неправильно, но система считает его распознанным правильно (такие случаи –не редкость), обнаружить такую ошибку будет очень сложно. В технологии Imagewriter визуально верифицируются все введенные тексты, а сравнивать исходный и введенный тексты очень удобно. Еще один метод верификации - верификация по правилам контроля корректности данных – используется в обоих технологиях и, в этом отношении, их возможности эквивалентны.

Сравнение стоимости подготовки, внедрения и эксплуатации апаратных средств технологий автоматического распознавания и Imagewriter показывает, что в них обоих в качестве рабочих мест операторов используются персональные компъютеры, объединенные локальной сетью, к одному из которых подключен оптический сканер, т.е. структура комплекса апаратных средств у них практически одинакова (различия могут быть только в количестве компъютеров, производительности и стоимости сканера).

Что касается стоимости лицензий на программное обеспечение, то тут следует учитывать различие областей применения этих технологий: ICR – для специально разработанных машиночитаемых бумажных форм, а Imagewriter – для любых бумажных форм. Поэтому сравнение возможно только по машиночитаемым формам. Для таких форм стоимость лицензии на ICR (например, FormReader), как минимум в 3-4 раза превышает стоимость лицензий Imagewriter при том, что производительность ICR выше в 2-2.5 раза.

Как отмечалось выше (см. п.2), в комплексах ICR обычно используются мощные промышленные сканеры (довольно дорогие), высока и стоимость лицензий на программное обеспечение ICR, поэтому для повышения эффективности использования этих средств комплексы ICR размещаются в специальных центрах сканирования, что, в свою очередь, вызывает необходимость решения проблемы доставки и архивации бумажных документов. Время, затрачиваемое на доставку, часто нивелирует эффект от использования ICR.

Технология Imagewriter относительна недорога, для ее использования не нужно создавать специальных центров с промышленными сканерами - подходит любой сканер. Это означает, что для ряда типов документов отпадает необходимость перевозки и сканирования в центрах сканирования и распознавания – они могут быть отсканированы там, где в этом возникает необходимость, а затем либо введены с помощью технологии Imagewriter, либо их образы будут отправлены в центры ввода по электронным каналам связи или на электронных носителях данных. Таким образом можно обеспечить избавление от бумажных носителей на самых ранних этапах обработки документов и обеспечить реальный переход к электронному документообороту.

5. Заключение

Проведенный сравнительный анализ технологий ввода данных с бумажных форм показывает, что каждая из рассмотренных технологий имеет и как достоинства, так и недостатки. Нельзя утверждать, что одна из рассмотренных технологий во всех отношениях лучше других. Очевидно, на выбор той или другой технологии в каждом конкретном случае будут влиять самые разнообразные факторы, возможно, даже не технические или финансовые. В то же время можно утверждать, что Imagewriter можно рассматривать как универсальную базовую платформу для создания средств ввода данных с любых бумажных форм, поскольку он может быть использован не только для ускорения ввода данных с образов бумажных форм, но и в качестве АРМ автоматизированного ввода с любых бумажных форм (без сканирования), а, при подключении модуля автоматического распознавания, может быть модифицирован в ICR.

Imagewriter позволяет для различных категорий пользователей предложить следующую последовательность наращивания производительности системы ввода данных в зависимости от финансовых возможностей:

- финансовые возможности очень ограничены - Imagewriter используется для быстрого создания АРМ‘а ввода данных и ввода данных непосредственно с бумажных форм (без сканирования);
- есть финансовая возможность закупки сканера и нескольких лицензий на Imagewriter - Imagewriter используется для разработки АРМ‘а ввода данных и ввода данных с образов отсканированных бумажных форм;
- есть финансовая возможность для закупки мощного и надежного сканера, подготовки и тиражирования машиночитаемых форм и закупки лицензий на модули автоматического распознавания на необходимое число рабочих мест - Imagewriter используется в качестве ICR.

Очевидна выгода использования Imagewriter для крупных корпоративных организаций, в которых собирается, вводится и обрабатывается много различных бумажных форм, например, отчетных. Среди всей совокупности типов бумажных форм могут быть выделены собираемые редко и в малых объёмах, собираемые более часто в относительно небольших объёмах и собираемые часто и в больших объёмах. Imagewriter позволяет обеспечить единый подход к созданию различных по производительности систем ввода данных на основе одной платформы. При появлении новых форм отпадает необходимость разработки нового АРМ‘а ввода - нужно лишь разработать новый шаблон для Imagewriter‘а и выбрать метод ввода. По мнению автора, технология Imagewriter должна найти широкое применение для обработки самых разных документов.

ВЫВОДЫ

1. В докладе проанализированы существующие технологии ввода данных и предложена технология ввода данных с любых бумажных форм по их графическим образам Imagewriter;
2. Проведен сравнительный анализ возможностей и эффективности различных технологий ввода данных;
3. Описан программный комплекс Imagewriter, обеспечивающий возможность ускорения ввода даных в 1.5-2 раза по сравнению с обычной технологией автоматизированного ввода данных;
4. Описаны возможности комплекса Imagewriter по быстрому созданию АРМ‘ов автоматизированного ввода с бумажных документов (без сканирования), АРМ‘ов автоматизированного ввода с образов документов и может использоваться в качестве базы для создания комплекса автоматического распознавания (при доукомплектовании модулем автоматического распознавания);
5. Показано, что технология Imagewriter позволяет обеспечить единый подход к созданию средств ввода данных с любых бумажных форм и выбор в каждом случае наиболее приемлимых по соотношению производительность/стоимость средств .

Литература

1. Мучник М.М. Технология сбора и отработки отчетных данных. - Корпоративные системы, №1, 2002, с.53-57.
2. Подольский И.Н. Печать на ПК слепым десятипальцевым методом.- Изд.2-е, доп.и перераб. - СПб:Наука и техника, 2002. – 80с.
3. Мучник М.М. Спосіб переносу символьних даних з графічних образів на носій інформації комп‘ютера цифровим кодом // Заявка № 2003054168 від 08.05.2003. - Рішення про видачу деклараційного патенту на винахід № 10833е від 21.05.2003. – Деклараційний патент на винахід №58428 А від 15.07.2003. Бюл. №7.