

# АВТОМАТИЗАЦИЯ ВВОДА ФОРМ

# Содержание

<b>Введение</b> .....	<b>3</b>
<b>Формы. Виды и области применения</b> .....	<b>3</b>
Что такое формы и где их применяют? .....	3
Как устроена форма? .....	4
Виды форм и способы их разметки .....	5
Что такое form processing? .....	6
Издержки ручного ввода .....	7
Автоматизация ввода форм .....	8
Основные принципы функционирования систем распознавания текста (OCR/ICR) .....	9
<b>Автоматизации ввода форм: шаг за шагом</b> .....	<b>10</b>
Когда требуется автоматизация? .....	10
Подготовка бланка формы .....	11
Разработка логической структуры формы .....	11
Выбор типа формы и разработка дизайна .....	11
Рисование бланка формы .....	12
Настройка системы на форму .....	13
Выбор сканера .....	14
Подготовка персонала .....	15
Цикл обработки данных .....	15
<b>Борьба за качество</b> .....	<b>17</b>
Что такое «качество ввода»? .....	17
Предварительная обработка изображений .....	17
Проверки по типам данных .....	18
Верификация .....	19
Проверка формата данных .....	20
Логический контроль .....	21
Обработка многостраничных форм .....	22
Спокойная работа оператора – еще одна гарантия качества! .....	22
<b>Как правильно организовать автоматизированный ввод документов</b> .....	<b>23</b>
Подходы к организации потокового ввода данных .....	23
Ввод данных во фронт-офисе .....	23
Ввод данных в бэк-офисе .....	24
Основные принципы потокового ввода данных .....	25
Пакетная обработка данных .....	25
Распределение функций операторов .....	25
Масштабируемость системы .....	25
Очередность заданий .....	25
Сохранение магистрали ввода .....	25
Проект по промышленному вводу форм .....	26
<b>Решение нетривиальных задач ввода с помощью технологий АBBYY</b> .....	<b>27</b>
Если система на поддерживает распознавание языка документа .....	27
Отделённое сканирование и обработка факсимильных документов .....	28
Распределённая верификация .....	28
Ввод «гибких форм» .....	29
Ввод данных с немашиночитаемых форм .....	29
<b>Заключение</b> .....	<b>31</b>
<b>Контакты</b> .....	<b>32</b>



## Введение

Вряд ли найдётся человек, которому хотя бы раз в жизни не доводилось заполнять бланки. Анкеты, счета и другие подобные документы давно существуют в различных областях человеческой деятельности. С другой стороны, сегодня для хранения и обработки информации повсеместно используются компьютеры и компьютерные сети. Неудивительно, что перенос информации с бумажных бланков в компьютерное хранилище

данных стал одной из самых актуальных задач в области документооборота.

В этом обзоре рассматривается процесс ввода данных с заполненных от руки бланков в память компьютера; анализируются особенности и преимущества автоматизированного ввода на примере используемой более чем в 30 странах мира системы ABBYY FormReader.

## Формы. Виды и области применения

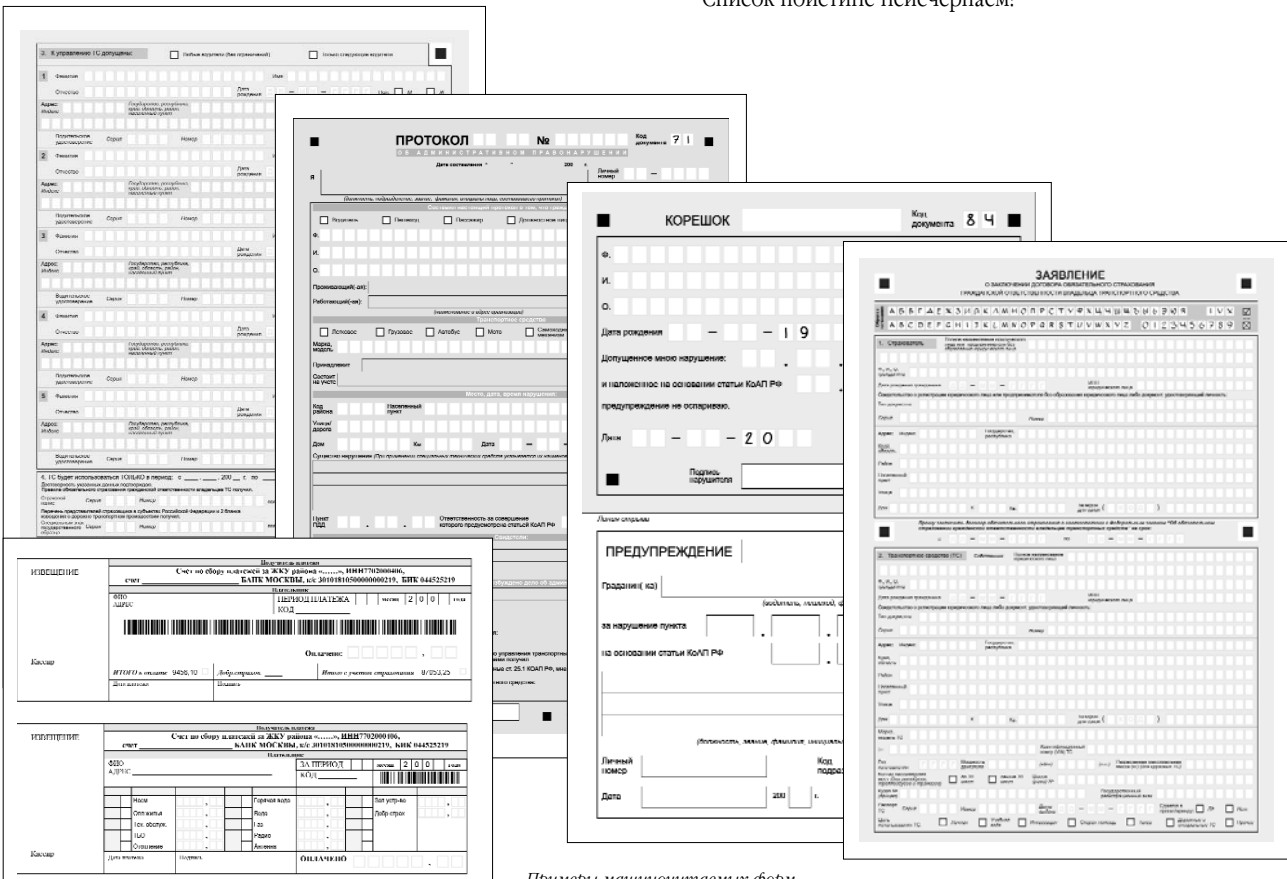
### Что такое формы и где их применяют?

Формой называется документ, имеющий фиксированную структуру и предназначенный для сбора определённой информации. Основные признаки формы – наличие чётко обособленных полей, пояснительных надписей, а также ряда служебных элементов, о которых мы поговорим дальше. В обиходе формы часто называют бланками.

Формы применяются повсюду, где необходимо опросить большое количество людей. Очень активно пользуются формами государственные учреждения, собирающие анкетные или иные данные, например, Министерство РФ по налогам и сборам или Пенсионный фонд РФ. В первом случае в виде заполняемых от руки форм составлены **налоговые декларации**, во втором – **анкеты пенсионного обеспечения**.

Столь же часто формы находят применение и в бизнесе. Страховые компании получают разнообразные виды документов-форм от своих клиентов: **заявление на получение полиса**, **заявление на возмещение ущерба** и т.д. Маркетинговые агентства вынуждены постоянно обрабатывать большое количество **опросных листов**. Образовательные учреждения проводят контроль за качеством знаний при помощи **тестов**. Весьма распространена процедура сбора данных при помощи форм в банковском деле – клиентами заполняются **заявления на получение кредитной карты** или **заявки на экспресс-кредитование**. А есть еще **торговые заказы**, отправляемые по почте, **рекламные купоны** на участие в розыгрыше призов, **формы медицинского обследования**, **квитанции** об оплате коммунальных услуг...

Список поистине неисчерпаем!



Примеры машиночитаемых форм.

При заполнении формы вся значимая информация заносится в поля – особым образом разграфлённые ячейки. Именно эта информация подлежит дальнейшей обработке. Формы, в которых определение положения полей и отделение данных от разметки может быть выполнено автоматически, программными средствами, называют машиночитаемыми. Вообще говоря, практически любая форма может быть приведена к виду машиночитаемой.

Форма может быть заполнена одним из следующих способов:

- ✓ от руки (такой способ заполнения называется рукопечатным: все символы пишутся раздельно, каждый символ занимает одно знакоместо);
- ✓ на пишущей машинке или матричном принтере;
- ✓ типографским способом;
- ✓ комбинированно, сочетанием вышеперечисленных способов.

## Как устроена форма?

При заполнении формы люди зачастую бывают невнимательными и неаккуратными. Во избежание ошибок формы составляются таким образом, чтобы сделать процесс заполнения интуитивно понятным. Для этого используют следующие **специальные элементы**, позволяющие ясно указывать заполняющему, какую информацию, куда и как следует вносить.

### Рассмотрим виды элементов.

- **Информационные поля** (entry fields). Предназначены для внесения собственно данных. Существует три вида информационных полей.
  - ✓ **Текстовые поля.** Каждое из них представляет собой группу знакомест, обычно с пояснительной надписью. Основное назначение знакомест – побудить заполняющего написать символы раздельно.
  - ✓ **Метки, или пункты** (checkboxes, checkmarks). Метка выглядит как одиночный замкнутый контур (квадрат, круг, многоугольник), снабжённый пояснительной надписью. Информация в такое поле вносится путём простановки условного знака («галочки», креста) внутри контура, либо путём его полного закрашивания.
  - ✓ **Группы меток.** Так называются несколько меток, расположенных рядом и объединённых по смыслу. Снабжаются пояснительной надписью возле каждой метки, а также общей пояснительной надписью, раскрывающей смысл вопроса. Как правило, метки внутри одной группы соответствуют взаимоисключающим вариантам ответа.
- **Сервисные поля.** В них располагаются реперные блоки (anchor points, reference points, definition blocks), используемые при распознавании. С их помощью программа определяет правильную ориентацию формы и выравнивает искажения при оцифровке изображения. Иногда сервисные поля служат для идентификации бланка при одновременной обработке нескольких различных форм. В качестве реперных блоков на формах для ввода с помощью ABBYY FormReader могут выступать следующие элементы:
  - ✓ сплошные квадраты чёрного цвета, углы и кресты;
  - ✓ сплошные линии: горизонтальные и вертикальные;
  - ✓ статический текст, то есть любая пояснительная надпись.

- **Идентификационные поля** (ID fields). Эти элементы предназначены для автоматической идентификации самого бланка формы. Реперный блок типа «квадрат», «угол» или «крест» обычно также может использоваться как идентифицирующий элемент. Но, как правило, для целей идентификации используют номер, нанесённый на форму в процессе изготовления бланка, само название формы или штрих-код.
- **Области для размещения графических изображений.** Используются для размещения графических (нераспознаваемых) объектов. В качестве примеров подобных объектов можно назвать блок введения подтверждающей записи, печать или штамп. При помощи ABBYY FormReader изображения из этих областей можно помещать в ODBC-совместимую базу данных в формате TIF, BMP, JPG, PCX или WMF.
- **Декоративные, необязательные элементы:** логотипы, колонтитулы и прочие элементы стилизации. При автоматизированном вводе информации зачастую используются для идентификации форм – анализируя текст в логотипе, программа может определить, от какой компании поступил данный документ (например, счет).



## Виды форм и способы их разметки

Выделяют два основных типа форм: структурированные и гибкие формы. К структурированным относят формы, поля которых не меняют размеры и взаимное расположение. Как найти данные на такой форме? Надо нарисовать её подобие – шаблон, который можно накладывать на поступающие изображения заполненных форм. Таким образом поля для распознавания как бы вырезаются из изображения, после чего буквы и цифры в этих полях распознаются.

Все бланки должны соответствовать образцу. На изображениях формы, полученной в результате сканирования, должны сохраниться реперные и идентификационные элементы.

На практике большинство форм не являются машиночитаемыми. Решение в большинстве случаев заключается в разработке и изготовлении новой, структурированной формы, отвечающей требованиям машиночитаемости.

По способу разметки выделяют **три основных типа форм**.

- **Цветная форма.** Все информационные поля на такой форме выполнены в виде белых прямоугольников на цветном поле. Чаще всего фон имеет серый, розово-оранжевый или зелёный оттенок. Цвет и насыщенность фона рекомендуется подбирать таким образом, чтобы его можно было легко удалить на этапе сканирования (drop-out colors). В идеальном случае после сканирования с формы должны исчезать все элементы, за исключением реперов и заполненных полей. Для подобной обработки используют либо специальные сканеры с цветной (красной или зелёной) лампой, либо особым образом выбранные настройки цветокоррекции в драйверах обычных сканеров. Наилучшее качество распознавания обеспечивается именно при использовании цветных форм.

Blank form with a color background (drop-out) for data entry. The form contains various fields for personal information and a grid of checkboxes for selection.

Цветная фоновая (drop-out) форма.

- **Растровая форма.** Информационные поля выглядят как белые прямоугольники на сером фоне. Фон состоит из растровых линий, состоящих из точек, которые расположены на одинаковом расстоянии друг от друга. После сканирования точки фона остаются на изображении. Однако технологии распознавания АBBYY позволяют удалять такие точки без потерь информации на заполненных полях.

Существует также обособленная разновидность растровой формы, где фон отсутствует. Границы информационных полей обозначаются растровыми линиями, состоящими из отдельных точек.

- **Чёрно-белая линейчатая форма.** Границы информационных полей на такой форме задаются обычными линиями, которые не исчезают при сканировании.

Form with black and white line markings for data entry. The form is titled 'Карта учета по всеобщей диспансеризации' and contains various fields for personal information and checkboxes for selection.

Форма с растровой разметкой границ полей.

Возможны следующие виды разметки линейчатой формы:

- текст по линии,
- текст в рамке,
- буквы в изолированных рамках,
- буквы в рамках,
- текст в «гребенке»,
- текст в рамке с «гребенкой».

Задача отделения содержания полей от их разметки для такой формы решается модулем распознавания. Основываясь на информации, указанной в атрибутах поля (тип разметки и количество ячеек), система находит вертикальные и горизонтальные полоски в разметке, затем она удаляет их, стараясь не повредить символы. Поскольку на форме может оказаться «мусор», также имеющий вид прямых линий, система «запомнит» все линии разметки, а лишнее будет удалять. Также отслеживаются точки соприкосновения элементов распознаваемого символа и «мусорных» линий. После очистки изображения производится распознавание символов.

Form with black and white line markings in a 'letter in frame' style. The form is titled 'Заявление о выдаче дубликата Страхового свидетельства' and contains various fields for personal information and checkboxes for selection.

Форма с черно-белой разметкой вида «буквы в рамках».

## Что такое form processing?

Ввод форм (form processing) – это перевод данных, содержащихся в информационных полях заполненных форм, в электронный вид; состоит из двух основных этапов:

- получение (захват) данных из формы (data capture);
- оцифровка и сохранение изображения исходной формы.

Как правило, процесс считается завершённым, когда все заполненные формы обработаны, а все данные введены, проверены и импортированы в формат используемой электронной

базы данных. При этом обычно требуется не только обеспечить высокое качество данных, но и минимизировать трудозатраты.

Существуют два основных метода ввода форм: вручную и с использованием средств автоматизации. В этой главе мы подробно рассмотрим особенности, преимущества и недостатки каждого из этих методов.

## Ввод форм вручную

Многие используют этот подход до сих пор, хотя он не оптимален, как с точки зрения надежности, так и с точки зрения трудозатрат. Почему? Судите сами.

Оценим, что понадобится сделать для подготовки к вводу форм ручным методом.

- Организовать рабочие места **операторов ввода**. Именно эта статья расходов оказывается самой весомой в затратной части бюджета. Средняя производительность труда квалифицированного оператора – до 200 насыщенных буквенными данными бланков в день. Требуется оснащение всех рабочих мест компьютерами, подключёнными к локальной сети.
- Организовать рабочие места **сортировщиков и контролёров входного потока**. В задачу контролёра, в частности, входит проверка комплектации многостраничных документов и общий надзор за процессом сортировки. Количество мест рассчитывают исходя из ожидаемых объёмов работы и средней производительности труда: до 1000 форм в день – для сортировщика и до 300 форм в день – для контролёра.
- Нельзя забывать также о рабочих местах **контролёров выходного потока**. В задачу этих сотрудников входит проверка качества данных, поступающих в электронном виде от операторов ввода, а также исправление ошибок, допущенных операторами.
- Кроме того, требуется привлечение **руководителя группы**, осуществляющего общий контроль и управление сотрудниками.

Оценим единовременные и регулярные затраты на отдел, который ежедневно вводит данные из 1000 форм. Для обеспечения такой производительности понадобится нанять пятерых операторов ввода, одного контролёра, а также одного руководителя группы. Для оснащения рабочих мест понадобятся соответственно семь столов, семь стульев, семь компьютеров

с мониторами, а также вспомогательная техника (сетевое оборудование, источники бесперебойного питания).

Статья расходов	Сумма расхода	Количество	Итого
	US\$		US\$
Компьютер	500	7	3500
Комплект мебели	500	7	3500
Сетевое и прочее оборудование	-	-	500
			7500

Таблица 1. Единовременные затраты при ручном вводе 1000 страниц в день.

Итак, размер разовых затрат – **7500** долларов США.

Теперь оценим объём ежемесячных затрат. Понятно, что не обойтись без аренды помещения, площадью около 50 кв.м. Допустим, арендная плата составляет 20 долларов в месяц за квадратный метр. Затраты на заработную плату операторов и контролёра положим равными 300 долларам в месяц, руководителя группы – 500 долларам.

Статья расходов	Сумма расхода	Количество	Итого
	US\$		US\$
Зплата оператора	300	5	1500
Зарплата контролера	300	1	300
Зарплата руководителя группы	500	1	500
Аренда офиса	20	50 кв. м	1000
			3300

Таблица 2. Ежемесячные расходы при ручном вводе 1000 страниц в день.

Заметим, что при этом не учитывались расходы на электроэнергию, телефонную связь, уборку помещений, затраты на резервный штат и т.д. В итоге даже при весьма скромной оценке получаем сумму порядка **3300** долларов в месяц.

## Издержки ручного ввода

Как следует из расчетов, приведенных выше, одновременные и ежемесячные затраты на ручной ввод, например, 1000 страниц в день составляют существенную сумму.

**Первый вывод: ручной ввод – это недешево.**

К сожалению, на этом проблемы, обычно сопутствующие ручному вводу, не заканчиваются. Как видим, требуется привлечение большого количества новых сотрудников, а также добавление дополнительного уровня управления. Очевидно, что подобную рабочую группу практически невозможно организовать в сжатые сроки. В самом деле, попробуйте быстро найти 8-10 человек, согласных на ваши условия. Зачислите их в штат, закупите технику и мебель. И не забудьте, что люди могут заболеть или даже совсем уволиться – не всякому подойдет такая утомительная работа.

А представьте, что клиент, заказывающий обработку форм, желает получить результат уже завтра (в крайнем случае – послезавтра), и станет понятно, что проблема цены – не единственная. Как за 2 дня набрать и усадить за работу 10 человек?

**Это второй вывод: систему ручного ввода нельзя организовать быстро.**

Заметим, что вне зависимости от количества сотрудников, производительность труда вашей рабочей группы не может быть увеличена оперативно, а сама группа оказывается практически немасштабируемой. Например, привлечение нескольких дополнительных операторов ввода бессмысленно, если не обеспечить их рабочими местами. Для организации этих мест надо арендовать дополнительную площадь. А есть ли она у арендодателя? Для контроля за информацией, вводимой новыми операторами, следует нанять дополнительных контролёров (им тоже нужны рабочие места). И так далее... Словом, любое расширение состава группы требует затрат времени

и средств, сравнимых с начальными затратами на организацию всей структуры.

**Третий вывод: систему ручного ввода нельзя быстро масштабировать.**

Существуют и другие проблемы. Наиболее существенны и практически неустранимы те из них, которые обусловлены человеческим фактором. Ручной ввод данных – занятие тяжелое: попробуйте набрать в текстовом редакторе хотя бы одну газетную статью. Поэтому даже опытные операторы допускают опечатки; причём к концу рабочего дня количество ошибок заметно возрастает. Часть из них устраняется контролёрами выходного потока, однако контролёры также подвержены усталости, поэтому качество данных в итоге существенно падает. Известно, что у профессиональных операторов ручного ввода неизбежно ухудшается зрение; уже через пару месяцев могут начаться непредвиденные сложности с персоналом.

Результаты очевидны – качество данных при ручном вводе оказывается низким. Человек, тем более уставший человек, не способен многие часы подряд тщательно и скрупулёзно выверять символ за символом. А значит, готовьтесь к проблемам с заказчиком: будет странно, если ему понравится кишачная ошибками база данных – плод труда целого отдела.

**Таким образом, делаем четвёртый и пятый выводы: людям не нравится такой труд; а вам не нравится качество их работы.**

Да, ручной ввод форм не является, мягко говоря, оптимальным вариантом. Особенно это верно для учреждений, проводящих сбор данных при помощи форм постоянно, а не периодически.

Ручной ввод форм.



## Автоматизация ввода форм

Альтернативный метод заключается в применении системы автоматизированного ввода данных. Рассмотрим особенности и основные стадии обработки форм с применением технологии ABBYY FormReader:

- пачку заполненных форм сканируют при помощи скоростного сканера (обычно применяют аппараты с производительностью не менее 10 страниц в минуту);
  - подавляющее большинство символов распознается автоматически;
  - символы, относительно которых сложилось несколько гипотез, автоматически передаются для проверки оператору системы ввода;
  - подтвержденная информация экспортируется в базу данных.
- Заметим, что на всех стадиях обработки требуется участие только одного человека – оператора ввода. Все операции, кроме укладки пачки форм в приёмный лоток сканера и проверки неуверенно распознанных символов, выполняются автоматически.

Рабочее место оператора ввода должно быть оборудовано сканером и одним компьютером, подключённым к локальной сети. Такое место может быть организовано в течение одного дня и не требует выделения больших дополнительных площадей.

Входной ручной сортировки поступающих бланков, а также ручной проверки комплектации многостраничных форм не требуется, поскольку система автоматизированного ввода способна самостоятельно идентифицировать формы и выбирать нужный шаблон распознавания.

Производительность труда одного оператора, использующего ABBYY FormReader 6.5 Desktop Edition, составляет от 1000 до 3000 страниц в день, в зависимости от сложности форм.

Давайте оценим разовые и ежемесячные затраты при использовании такой системы на одном рабочем месте из расчета тех же 1000 страниц в день.

Статья расходов	Сумма расхода US\$	Количество	Итого US\$
Компьютер	500	1	500
Сканер	1500	1	1500
Комплект мебели	500	1	500
Лицензия ПО	1600	1	1600
Внедрение ПО	800	1	800
			4900

Таблица 3. Единовременные затраты для автоматизированного ввода 1000 страниц в день.

Статья расходов	Сумма расхода US\$	Количество	Итого US\$
Затраты на основного оператора	500	1 чел.	500
Затраты на резервного оператора	300	1 чел.	300
Аренда офиса	20	10 кв. м	200
Т/О сканера (в расчете за месяц)	-	-	20
			1020

Таблица 4. Ежемесячные расходы для автоматизированного ввода 1000 страниц в день.

А теперь сравним полученные результаты.

	Ручной ввод US\$	Ввод с FormReader US\$	Экономия US\$
Разовые затраты	7500	4900	2600
Ежемесячные затраты	3300	1020	2280

Таблица 5. Экономия средств при необходимости ввода 1000 страниц в день.

Цифры говорят сами за себя. Но самое важное то, что выбор автоматизированной системы ввода раз и навсегда **решает все пять проблем**, описанных выше!

Система автоматизированного ввода может быть неограниченно масштабирована, для этого потребуются приобрести нужное количество дополнительных копий ABBYY FormReader и привлечь к работе операторов, обучение которых займет несколько часов. **Вы знаете другой способ десятикратно увеличить производительность системы ввода за 1 день?**

И, конечно, резко возрастает качество. Как показывает практика, качество данных при автоматизированном вводе форм оказывается на несколько порядков выше. Причины этого очевидны: влияние человеческого фактора сведено к нулю. Основной объём работы выполняется компьютером, который не устает и никогда не допускает опечаток. Кроме того, система ABBYY FormReader снабжена набором встроенных правил контроля, которые существенно повышают общую надежность системы и, следовательно, качество данных.

Автоматизированный ввод форм.



## Основные принципы функционирования систем распознавания текста (OCR/ICR)

Выделяют два основных класса систем оптического распознавания символов: OCR (optical character recognition) и ICR (intelligent character recognition). OCR-системы распознают печатные символы, нанесенные на бумагу типографским способом, при помощи принтера, плоттера или пишущей машинки. ICR-системы обрабатывают документы, заполненные печатными буквами и цифрами от руки, или, иначе говоря, распознают рукопечатные символы.

Рассмотрим, чем различаются принципы действия этих систем. OCR-система в процессе анализа выделяет на изображении блоки (текст, таблицы, иллюстрации), затем последовательно разделяет блоки на всё менее крупные объекты: абзацы, строки, слова, символы. Последние обрабатываются программными механизмами, осуществляющими собственно распознавание; эти механизмы называют классификаторами. Затем распознанные символы «собираются» в слова, слова – в строки, и так далее, вплоть до синтеза полного электронного аналога исходного документа.

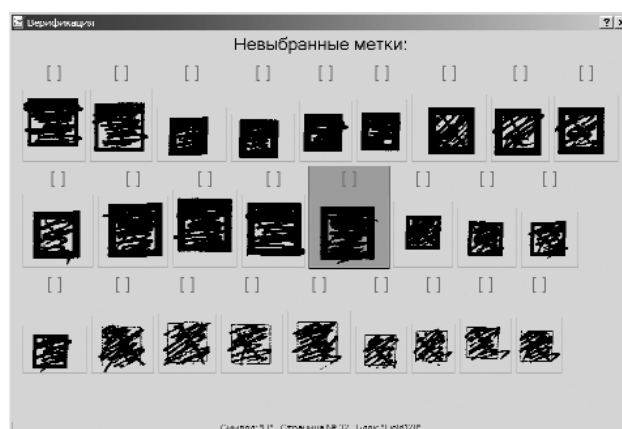
ICR-система, нацеленная в первую очередь на обработку форм, функционирует иначе. На исходном изображении выделяются области, в которых должна содержаться смысловая информация, и затем именно эти фрагменты подвергаются дальнейшей обработке, в том числе и при помощи классификаторов. Иначе говоря, ICR-система не пытается построить точную электронную модель документа, а лишь извлекает информацию из четко ограниченных областей. Эта информация передаётся в систему хранения.

К ICR-системам предъявляется также требование по распознаванию специальных объектов – меток (mark sense). Ведь использование в формах меток, как отмечалось, позволяет упростить заполнение форм и значительно повысить качество ввода, вплоть до 99,9%. ABBYY FormReader 6.5 способен распознавать метки произвольного вида (квадратные, круглые и т.п.). Используемая для этого технология OMR (Optical Mark Recognition) заключается в следующем.

При построении шаблона формы оператор отмечает область вокруг метки, подлежащую анализу. Система анализирует и сохраняет информацию о распределении чёрного цвета в указанной окрестности метки. (Естественно, при этом необходимо, чтобы метка на исходной форме не была закрашена.

Впрочем, система ABBYY FormReader 6.5 в состоянии обрабатывать метки типа «пустое место» и «прямоугольная рамка» даже в том случае, если метка на исходном изображении закрашена.) Далее, при распознавании очередной формы система анализирует распределение чёрного цвета. Если уровень затемнения отчетливо превышен, выносится решение о том, что данное знакоместо отмечено.

Подобная технология позволяет ABBYY FormReader 6.0 уверенно распознавать не только обычные пометки («галочки», крестики), но и выделять полностью закрашенные метки в том случае, если последние имеют вид прямоугольной рамки или поля без границы.



Верификация полностью закрашенных меток в ABBYY FormReader 6.0 Desktop Edition

Эта способность ABBYY FormReader находит очень важное применение. Представьте, что человек допустил ошибку при заполнении формы, сразу это понял, но... «галочка» уже стоит! Что делать? Брать новый бланк, заполнять заново? Есть более простое и даже остроумное решение. Заполняющий полностью закрашивает метку, выделенную по ошибке. ABBYY FormReader определит её как отмеченную по ошибке, то есть неотмеченную. Подобный метод может применяться и для текстовых полей.

## Автоматизация ввода форм: шаг за шагом

### Когда требуется автоматизация?

Обстоятельства, в которых возникает потребность в автоматизированном вводе форм, могут быть самыми различными. Опишем несколько характерных случаев.

- **Обработка форм не является для компании профильной деятельностью.** Например, это производственная или торговая компания. Как правило, в структуре компании даже отсутствует подразделение, специализирующееся на вводе форм. Обработка входящих документов (например, заявок на поставку продукции) в таком случае осуществляется силами секретарей приёмной. Пока объём поступающих форм исчисляется единицами или десятками, особых проблем не возникает. Но когда количество заявок превышает сотню в день, руководству приходится нанимать дополнительных сотрудников. В противном случае в приёмной возникают очереди, да и сами сотрудники вынуждены отвлекаться от общения с клиентами, выполнять не свойственные им задачи.

Решение – автоматизация ввода входящих данных, например при помощи ABBYY FormReader Desktop Edition. Система может быть размещена на одном рабочем месте, практически не требует расширения штата и дополнительного обслуживания.

- **Обработка анкет – один из основных бизнес-процессов в компании.** Пример – маркетинговое агентство, реализующее сбор и обработку данных. Задача обработки характеризуется прежде всего большим объёмом поступающей информации. Агентство может нуждаться в обработке до 10 тыс. страниц в день и даже больше, причём ввод данных здесь – часть основного технологического процесса. И требования к системе, выбираемой для автоматизации ввода, имеют свои особенности.

Во-первых, объём поступающей информации зависит от пожеланий заказчика того или иного маркетингового исследования. Понятно, что этот объём может сильно изменяться в зависимости от условий проведения очередного исследования. Поэтому особенно важно обеспечить хорошую **масштабируемость** – возможность быстрого расширения системы автоматизированного ввода данных.

Во-вторых, в силу специфики деятельности маркетингового агентства, вложение средств в автоматизацию обработки данных представляет собой вложение в основные средства производства. Поэтому для технико-экономичес-

кого обоснования вложений необходимо, чтобы **возврат на инвестиции (ROI)** во внедрение системы был заранее известен и имел приемлемое значение.

В-третьих, анкеты могут существенно меняться от проекта к проекту. Поэтому маркетинговому агентству неплохо бы иметь в своем распоряжении удобное **средство для рисования новых форм**.

Такая система автоматизированного ввода форм, как ABBYY FormReader Enterprise Edition, удовлетворяет всем этим условиям. Масштабирование системы может осуществляться неограниченно – как простым увеличением количества станций, так и за счёт организации распределённой обработки данных.

- **Перевод архива в электронный вид.** Чаще всего эта задача возникает одновременно. Однако объём подлежащей оцифровке информации при этом весьма велик – в «бумажном» виде архив обычно занимает несколько комнат, целиком заставленных стеллажами. В то же время владельцы архива обычно не располагают финансовыми и организационными возможностями для найма дополнительных сотрудников.

В данном случае не столь важно время, которое может потребоваться для организации автоматизированной системы ввода данных. Самое главное – простота решения. Оптимальным считается вариант, который может быть реализован без привлечения большого количества людей и серьёзных вложений. В частности, таким вариантом является организация одного рабочего места оператора на базе системы автоматизированного ввода форм ABBYY FormReader.

Для такого случая у компании ABBYY есть специальная система лицензирования ABBYY FormReader – так называемая модель Page Count. Эта схема лицензирования предполагает, что пользователь приобретает возможность ввода ограниченного объема страниц. Соответствующее ограничение задается в ABBYY FormReader. Исчерпав его, ABBYY FormReader переходит в нерабочее состояние. Когда нужно один раз ввести известное количество страниц, такой подход оказывается наиболее удобным и оправдывается с финансовой точки зрения.

Возможно на вашем предприятии существуют похожие задачи. Но как приступить к решению? С чего начать? Что предпринять в сложных ситуациях?

## Подготовка бланка формы

Работа начинается с подготовки бланка, который будут заполнять опрашиваемые. Очень важно создать форму, удобную как для заполнения, так и для обработки. Ошибки, допущенные при разработке бланка формы, могут катастрофически снизить эффективность всего процесса. Поэтому на всех стадиях подготовки следует строго придерживаться рекомендаций, которые исходят от поставщика системы автоматизированного ввода.

Изготовление бланка формы состоит из трёх основных стадий: разработка логической структуры, разработка её дизайна и рисование бланка формы. Рассмотрим все эти стадии подробнее.

### Разработка логической структуры формы

Чем лучше продумана структура, тем проще в дальнейшем будет заполнять и обрабатывать форму. Определите, какие именно данные понадобятся вводить, составьте и согласуйте со всеми заинтересованными лицами список информационных полей.

Затем следует определить такие важные параметры, как формат и количество листов формы. Обратите внимание: смена формата впоследствии может привести к необходимости вносить существенные изменения и в бланки формы, и в настройки системы! Именно поэтому советуем сразу рисовать все эскизы на листах выбранного формата, чтобы не столкнуться с нехваткой места для элементов формы.

**Идентифицирующее поле для многостраничных форм (ID field).** Если выяснилось, что нужна многостраничная форма, сразу продумайте, каким образом избежать путаницы – как между страницами, так и между формами? Обычно в таких случаях каждую страницу снабжают специальным идентифицирующим полем. Данные, внесенные в это поле, должны быть одинаковы на всех страницах формы. Какое именно поле для этого выбрать, зависит от предметной области. Например, это может быть ИНН физического лица, номер страхового картонки, БИК банка, номер карточки социального страхования, учетный номер клиента, шифр проекта и т.д.

**Простые и составные поля.** Продумывая компоновку формы, старайтесь составлять поля как можно проще. Дело в том, что количество ошибок заполнения и ошибок распознавания на простых полях оказывается куда меньше. Чем точнее можно задать множество слов (или символов), которые могут встретиться в данном поле, тем выше будет качество распознавания. Весьма желательно разделять на несколько составляющих такие поля, как «ФИО», «дата», «телефон» («код города» + «номер»), «адрес» («страна» + «город» + «улица» + ...).

Свободное место на форме всегда в дефиците, поэтому если известна максимальная длина поля, то под это поле надо вводить **ровно необходимое число позиций**. Это поможет дисциплинировать заполняющего, а сам процесс заполнения сделать для него более удобным. Примеры текстовых полей с известным числом позиций: ИНН, почтовый индекс, почтовое сокращение для штата в США, номер телефона для локально распространяемых анкет, знаки стандартизации, сокращенное название валюты.

**Длина полей.** Длина слов в таких полях, как «название улицы», «фамилия» или «e-mail» может быть почти произвольной, поэтому для них количество знаменослов следует выбирать с запасом. Если есть большая вероятность того, что длины одной строки будет недостаточно, отведите под это поле две или более

строки. Система позволяет объединить их в одно поле прямо в процессе распознавания, так что на качество это не отразится.

**Разделители.** Желательно сделать форму такой, чтобы заполняющий вносил в неё только значимую информацию. Например, в поле «дата» желательно сразу расставить символы-разделители (точки, тире или наклонные черты). Пусть заполняющий впишет только цифры – это заметно повысит точность распознавания. Другие примеры: можно сразу проставить на форме дефис для SSN или ГОСТА, первые три цифры года.

**Метки (checkmarks).** В тех случаях, когда заранее известны все возможные варианты ответа, вместо текстовых полей лучше использовать метки. Алгоритмы OMR (Optical Mark Recognition) позволяют определять наличие рукописных меток с очень высокой вероятностью, что гораздо выше показателей для распознавания рукопечатного текста. Поэтому при первой возможности старайтесь заменять текстовые поля метками или группами меток. Например, вместо текстового поля «семейное положение», в которое заполняющий сможет вписать произвольное слово («женат», «замужем», «разведен», «холост» и т.д.), рекомендуется создать группу из 3-х меток («не состою в браке», «состою в браке», «разведен (а)»).

**Подписи и фотографии.** Если необходимо разместить на форме такие поля, как «подпись», «печать», «фотография», «отпечатки пальцев», постарайтесь отвести для них достаточно места. Тогда количество помарок при заполнении будет меньше, а качество распознавания – выше. Обратите внимание: когда ставят печать или приклеивают фотографию, на обратной стороне листа зачастую проступают пятна; поэтому необходимо убедиться в том, что это не помешает обработке информации с другой стороны формы.

### Выбор типа формы и разработка дизайна

Отделение содержимого полей от разметки – одна из главных проблем распознавания текста. И то, насколько успешно она решается, во многом зависит от правильного выбора типа формы. Внесенная в поля информация должна быть корректно отделена от прочих элементов: границ полей, фона, служебных и пояснительных надписей. Напомним, что наиболее удобными с данной точки зрения являются так называемые цветные фоновые формы (drop-out forms). Фон дисциплинирует заполняющего, поскольку задаёт границы полей и отдельные знаменослов, а проблем для распознавания не создает, так как отсеивается при сканировании. Простой критерий выбора может быть сформулирован так: используйте серые фоновые формы во всех проектах, где невозможна типографская печать цветных фоновых форм.

Занимаясь разработкой дизайна, непременно обратите внимание на правильный выбор и расстановку реперных и идентификационных элементов формы – тогда автоматический ввод данных будет максимально эффективен.

**Что такое реперы?** Для точного наложения шаблона система должна иметь возможность «опереться» на некие элементы. Их принято называть реперными блоками или реперами. Благодаря им программа может отслеживать линейные искажения и сдвиги изображения, а также определять расположение полей. Такие элементы принято называть также «якорями» (anchors). Примеры реперных полей: черные квадраты, углы, кресты, не исчезающие при сканировании надписи, линии. Специалисты компании ABBYY рекомендуют размещать три

или четыре черных квадрата по углам страницы. Их наличие позволяет программе точно и быстро накладывать шаблон формы, вводить в едином потоке формы, напечатанные на разных принтерах, и формы, переданные по факсу.

**Что такое идентификатор?** Это элемент, который после сканирования сохраняется на изображении формы и используется для наложения шаблона на форму. В случае одновременной обработки нескольких форм в одном потоке необходимо на каждой странице формы предусмотреть уникальный элемент, указывающий на принадлежность страницы к той или иной форме. В качестве идентификаторов рекомендуется использовать штрих-код, название формы или дополнительный черный квадрат.

## Рисование бланка формы

Как правило, когда уже продумана логическая структура формы и нужно переходить к разработке дизайна – возникает закономерный вопрос: какие программные инструменты лучше всего использовать? Мы предлагаем краткий обзор соответствующих программ.

Если в штате организации есть дизайнер, умеющий работать с CorelDRAW или Adobe Illustrator, лучше всего прибегнуть к его услугам. Эти программы наиболее удобны, но у них есть своя специфика: оба графических пакета – «тяжёлые», профессиональные инструменты, к тому же весьма недешёвые. Работа с ними может оказаться непосильной для новичка, а освоение всех возможностей займёт много времени.

Пакет Microsoft Visio более распространён и менее сложен. Хотя он предназначен для рисования графиков и схем, с его помощью можно при желании делать неплохие формы. Проще всего использовать для этого так называемые галереи трафаретов. Подобную галерею, содержащую элементы форм: квадраты, поля для ввода и т.д., – можно получить, например, в компании ABBYY. Так можно создавать вполне профессио-

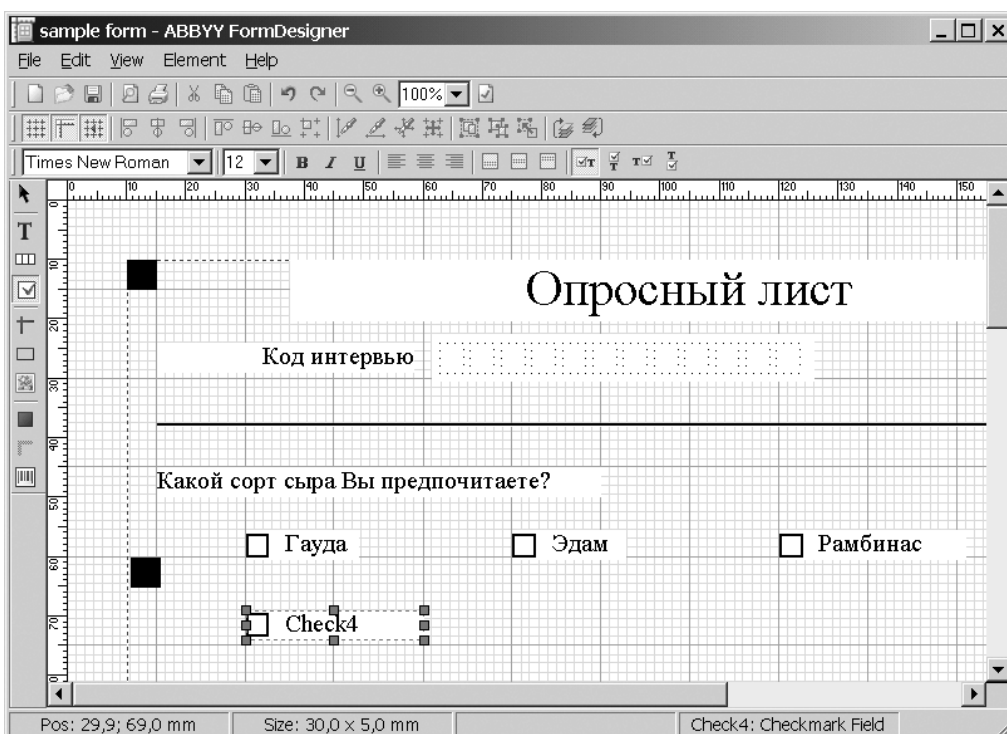
нальные серые фоновые формы и печатать их потом на лазерном принтере.

В самом крайнем случае – если нет ничего более подходящего – можно создавать несложные формы и при помощи общеизвестного текстового редактора Microsoft Word. Поскольку система предназначена для других целей, рисование формы при помощи текстового процессора будет не самым легким и приятным процессом...

**Впрочем, есть способ намного удобнее.** Программа **FormDesigner** из комплекта поставки ABBYY FormReader специально создана для рисования форм. Этот простой и удобный инструмент позволит быстро и безошибочно изготовить форму любой сложности.

Формы содержат определенные типовые элементы: название бланка, черные квадраты, текстовые поля, состоящие из названия и ячеек для ввода, метки и т.д. Удобно, когда все эти элементы заранее нарисованы, а их параметрами, такими как размеры или вид рамки, легко управлять, задавая нужные значения. Программа FormDesigner позволяет создавать и редактировать формы в соответствии с принципами WYSIWYG. Всё, что требуется от дизайнера, – просто переносить при помощи мыши стандартные элементы из галереи на бланк формы. Это даёт возможность не тратить время на поиски специальных графических примитивов. Когда работа над формой завершена, создается файл формата XFD, в котором будет храниться автоматически созданная разметка шаблона формы. В дальнейшем, при настройке программы на работу с этой формой, достаточно лишь импортировать готовый xfd-файл, указать нужные атрибуты уже размеченных полей, задать правила проверок и скорректировать, если нужно, набор реперных блоков.

Когда бланк готов, нужно настроить систему на работу с полученной формой.



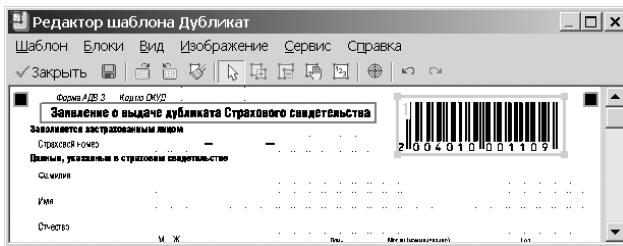
Рисование бланка формы при помощи ABBYY FormDesigner

## Настройка системы на форму

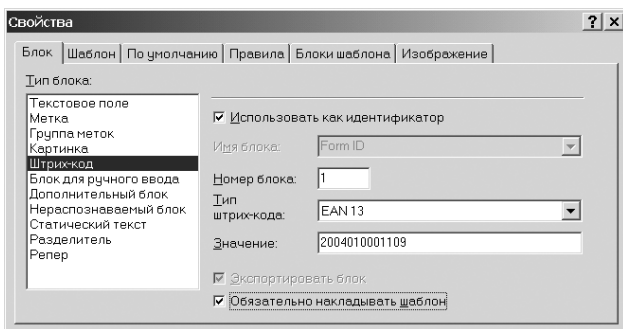
Смысл этих действий – «объяснить» системе, как именно следует воспринимать те или иные элементы формы, в каких областях искать поля для ввода данных, какие «подсказки» использовать при распознавании. Это не менее важный этап, чем предыдущий.

**Изготовление шаблона формы.** Здесь мы опишем полную, пошаговую последовательность изготовления шаблона.

1. Процесс начинается с получения изображения незаполненной формы. Чаще всего для этого просто сканируют готовую форму; впрочем, ABBYY FormReader допускает использование изображения, полученного ранее. Если бланк формы подготовлен с использованием ABBYY FormDesigner, процесс можно существенно облегчить. Для этого достаточно импортировать в ABBYY FormReader шаблон в формате XFD, созданного при помощи ABBYY FormDesigner. В таком шаблоне на изображении формы уже будут содержаться размеченные блоки. Если же изображение формы получено путем сканирования существующей бумажной формы, то необходимо провести все нижеуказанные действия.
2. Определение реперов и идентификаторов. Сами по себе эти блоки могут быть выделены на изображении автоматически либо вручную. Идентификационные блоки иногда могут представлять собой статический текст либо штрих-код.

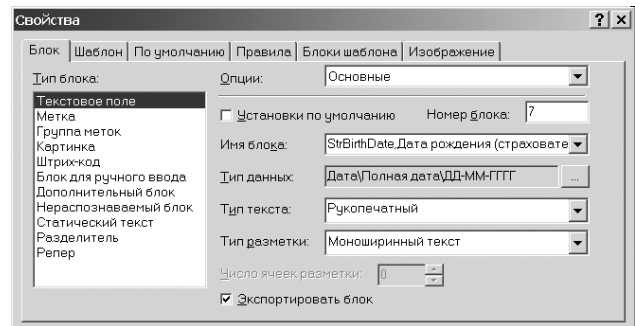


Определение реперов и идентификаторов в редакторе шаблона формы.



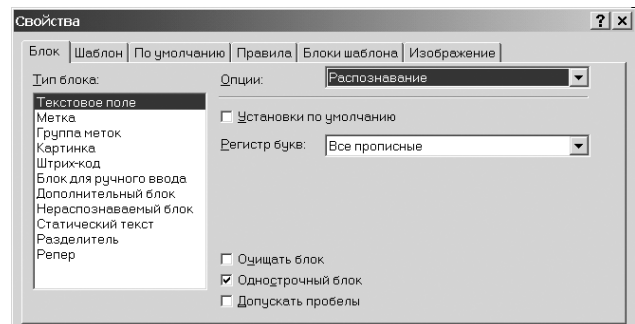
Определение штрих-кода в качестве идентификатора формы.

3. Проверка наложения шаблона – совершенно необходимый шаг. Удостоверившись, что созданные реперы и идентификаторы позволяют точно совместить нужный шаблон с изображением формы, переходим к следующей стадии.
4. Геометрическая разметка полей для распознавания. Создать распознаваемый блок можно буквально одним движением мышки.
5. Затем опишем свойства созданных полей, такие как «имя блока», «тип данных», «тип текста», «тип разметки» и т.д. Сначала, однако, рекомендуем проанализировать шаблон и определить параметры, общие для большинства блоков, и принять их как «параметры по умолчанию».

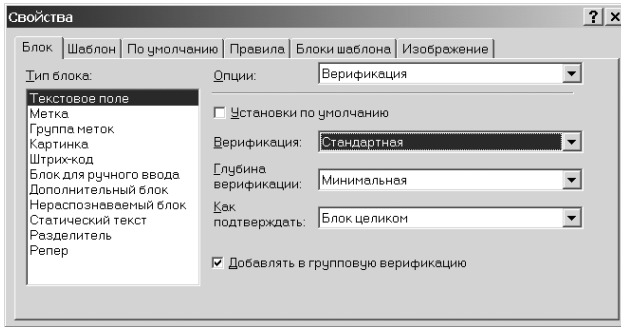


Определение основных атрибутов поля в редакторе шаблона формы.

6. Необходимо определить правила проверки. Правила – это некоторые условия, накладываемые на данные в полях и автоматически проверяемые программой. С помощью правил можно проверить формат распознанных данных и привести их к стандартному виду (например, правило проверки и нормализации даты), проверить информацию по базам данных и спискам допустимых значений, можно проверить соответствуют ли друг другу данные из нескольких блоков (например, что сумма цифрами равняется сумме прописью).



Определение опций поля, связанных с распознаванием.



Определение опций поля, связанных с верификацией.

7. Задание нужных параметров опций в разделах «Распознавание» и «Верификация» также может существенно повлиять на качество распознавания.
8. Если данные предполагается сохранять в базе данных, используя для этого ODBC-соединение, настройка такого экспорта также должна быть проведена в редакторе шаблона формы.

До того как начать потоковый ввод, необходимо определиться также с выбором сканирующего устройства.

## Выбор сканера

От параметров выбранного сканера будет зависеть и скорость и качество обработки данных. Следует сразу отметить, что при большом количестве форм (свыше 100 ежедневно) обычные планшетные сканеры неприменимы. Эти устройства широко распространены в офисах, они неплохо справляются с оцифровкой фотографий и обычных документов, но для потокового ввода непригодны: у них невысокое быстродействие и относительно небольшой ресурс. **Только представьте, во что превратится крышка планшетного сканера после сканирования нескольких тысяч страниц!**

Для полноценного, быстрого и качественного ввода большого количества форм нужен специальный аппарат, предназначенный для ввода большого числа, как правило, однотипных документов.

На что нужно в первую очередь обратить внимание при выборе сканера для автоматизированного ввода форм?

- **Формат бумаги.** Чаще всего для ввода форм используются устройства, способные сканировать листы формата A3, A4 и A5.
- **Разрешение изображений.** Для ввода форм требуется разрешение 200–300 dpi (dots per inch – точек на дюйм) и все сканеры поддерживают такие режимы. Сканирование с более высоким разрешением приводит к неременному замедлению, а скорость сканирования может быть одним из самых критичных параметров при потоковом вводе форм.
- **Двустороннее сканирование.** Для многих проектов необходимо применять сканеры, которые могут осуществлять как одностороннее, так и двустороннее сканирование в черно-белом или цветном режимах. Последний режим необходимо использовать, например, при очистке изображения от цветных печатей и сохранении цветных фотографий с анкет.
- **Наличие устройства для автоматической подачи бумаги – автоподатчика (ADF, Automatic Document Feeder).** Это устройство, позволяющее загружать формы в сканер пачками обычно по 25, 50 или 100 документов, необходимо практически в любом случае, иначе работа оператора ввода будет на 90% состоять из манипуляций с бумагой и сканером.

- **Производительность.** Часто скорость работы всей системы автоматизированного ввода зависит именно от быстродействия выбранного сканера. Выделяют три основные категории сканеров: офисный малой производительности, офисный средней производительности и высокопроизводительные промышленные сканеры. Их ежедневная нагрузка: 500 листов – для сканеров малой производительности, более 20 тысяч листов – для промышленных высокопроизводительных моделей.
- **Контроль двойного захвата листа.** Захват протяжным механизмом сканера более чем одного листа бумаги может привести к тому, что какая-то форма вообще не будет обработана. Для предотвращения подобной ситуации во многих сканерах реализованы системы контроля: при помощи взвешивания захваченной бумаги, замера толщины бумаги или контроля светового потока, проходящего через сканируемую бумагу.

Однако эти способы неприменимы, если поток форм неоднороден, т. е. если вводятся формы различных видов (на разных форматах бумаги, разного цвета, плотности и т.д.). Поэтому наибольшее распространение получают системы контроля на базе ультразвуковых датчиков, которые следят за тем, чтобы отраженный сигнал приходил не более чем от одной поверхности, то есть от одного листа бумаги.

- **Наличие специальных возможностей.** Некоторые аппараты оборудованы вспомогательными возможностями, которые могут оказаться очень полезными. Среди них:
  - принтер для надпечатки на одной из сторон отсканированного документа (endorser) специального индекса для идентификации документа в дальнейшем;
  - аппаратный модуль для улучшения качества получаемого изображения;
  - аппаратный модуль компрессии изображений;
  - цветные лампы подсветки для удаления определенного цвета с так называемых «фоновых» (drop-out) цветов, обычно красного, реже – синего или зеленого;
  - кеширование изображений с использованием собственной памяти сканера, что также повышает быстродействие системы.

## Подготовка персонала

Для работы с ABBYY FormReader практически не требуется специально обученного персонала. Обычно привлекаются операторы, которые выполняют потоковый ввод форм, и администратор комплекса, который занимается настройками и выполняет контрольные функции.

- Существуют две разновидности операторской работы:
  - операции выполняются на одном компьютере; в обязанности оператора входит загрузка форм в сканер, контроль над процессом сканирования и распознавания, верификация данных;
  - при использовании ABBYY FormReader Enterprise Edition операторов несколько; каждый из них выполняет только одну из следующих функций: сканирование, проверка сборки многостраничных документов, верификация или экспорт данных.
- Администратор комплекса по вводу форм выполняет настройку системы. В тех случаях, когда к запуску готовится

комплекс на базе версии Enterprise Edition, от администратора требуется не только подготовить шаблоны форм, но также развернуть систему, распределить роли операторов, задать описания многостраничных документов и т.д. Затем, в процессе работы комплекса, в его обязанности будет входить мониторинг информационных потоков.

Обучение операторов и администратора\* занимает от нескольких часов до двух-трех дней, состоящих из практической работы, – за это время приобретаются все необходимые навыки.

\* Программа обучения администратора комплекса :

- 1) администрирование оборота форм в компании;
- 2) создание новых форм;
- 3) подготовка шаблонов форм;
- 4) инсталляция продуктов ABBYY, в том числе сетевых;
- 5) настройка опций сканирования, распознавания, верификации;
- 6) администрирование ролей пользователей системы;
- 7) настройка правил проверки и сборки многостраничных документов;
- 8) мониторинг работ и формирование отчетов в системе.

## Цикл обработки данных

Для того чтобы дать представление об особенностях работы системы автоматизированного ввода форм ABBYY FormReader, опишем в общих чертах цикл обработки данных.

1. **Открытие пакета.** Пакетом называется множество однотипных документов, каждый из которых к концу обработки представлен как изображение и как набор упорядоченных, готовых к экспорту данных (значений полей). Открытие пакета (нового или созданного ранее) подразумевает приведение системы в состояние готовности к работе. Выполняется оператором либо автоматически.
2. **Добавление в пакет изображений.** Изображения подлежащих обработке форм можно добавлять в пакет одним из трёх способов:
  - сканированием бумажных форм;
  - добавлением в пакет изображений из созданных ранее графических файлов;
  - «перетаскиванием» (drag-and-drop) значка файла, например, используя обозреватель (browser) Microsoft Windows.
3. **Распознавание.** Этот процесс выполняется автоматически и представляет собой перевод имеющегося на изображении текста в электронный вид. Сначала выполняется автоматическое наложение шаблона, после чего на изображении выделяются блоки, предназначенные для распознавания. Затем изображение каждого блока распознаётся, то есть преобразуется в собственно текст.

4. **Проверка результатов.** После распознавания всех изображений пакета часть страниц может содержать неуверенно распознанные символы. Такие страницы поступают на ручную проверку (верификацию). Во время верификации оператор либо подтверждает правильность символов, либо исправляет те из них, которые были распознаны неверно. Аналогично исправляются ошибки, обнаруженные правилами контроля. Система помечает страницы, на которых правила не выполняются корректно, специальными флагами ошибки или предупреждения.
5. **Экспорт.** Проверенные данные переносятся в файл указанного формата либо в базу данных. Для этого оператору достаточно нажать на кнопку «Экспорт»

Как видим, участие оператора в процессе ввода данных минимально. А главное, у оператора практически нет свободы выбора (одного из источника ошибок): система проводит оператора по всем стадиям от сканирования до экспорта. Наряду с прочими преимуществами, автоматизированный ввод данных из форм позволяет добиться более высокого, чем при ручном вводе, качества данных. Это достигается за счет применения специальных средств, о которых будет рассказано дальше.

На этой странице приводится схема работы системы потокового ввода форм ABBYY FormReader 6.5 Enterprise Edition.

- Система имеет один входной и один выходной поток данных.
- Каждый оператор выполняет только одну операцию, например, следит за сканированием и регистрацией пакетов.
- Операторы работают по конвейерному принципу.
- Если каких-либо операторов недостаточно, например операторов верификации, их количество можно увеличить.
- В системе осуществляется централизованное хранение данных и параметров настройки.
- Защита комплекса происходит централизованно, с помощью электронного ключа, установленного на сервере.

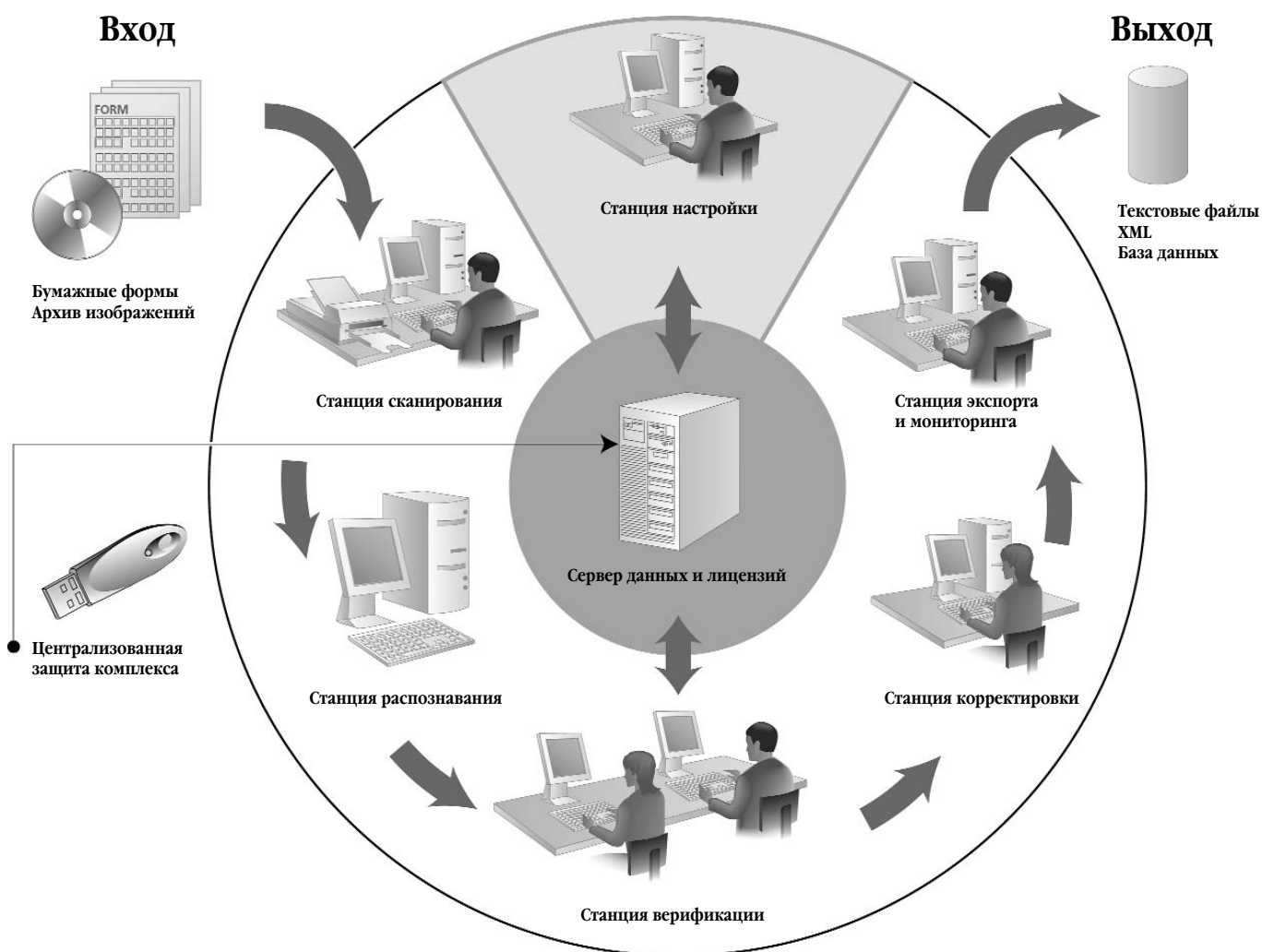


Схема обработки данных в системе потокового ввода форм ABBYY FormReader 6.5 Enterprise Edition.

# Борьба за качество

## Что такое «качество ввода»?

Мы неоднократно упоминали интуитивно понятный термин «качество ввода», пришло время дать ему определение.

Под качеством ввода понимается степень соответствия информации, поступающей в систему хранения (target system), той, что была внесена при заполнении. Чем точнее сведения в базе данных соответствуют тем, что содержатся на заполненных формах, тем выше качество данных.

Качество ввода – один из важнейших параметров, характеризующих систему автоматизированного ввода форм. Вот несколько основных факторов, снижающих качество.

- **Неаккуратное заполнение формы.** Если заполняющий допустил помарки или исправления, либо просто написал некоторые буквы слитно, вероятность ошибок распознавания возрастает. Способ противодействия очевиден – при разработке форм чётко обозначать знакоместа для каждого знака и разбивать каждое составное поле на ряд простых. Если работа по дизайну формы выполнена с учетом рекомендаций, описанных в разделе «Разработка логической структуры формы», влияние качества заполнения на общее качество ввода будет минимальным.
- **Опечатки.** При вводе форм вручную удельный вес этого фактора очень велик. Операторы неизбежно устают и количество допускаемых ими опечаток, относительно небольшое в начале рабочего дня, резко возрастает к вечеру. Единственное радикальное средство борьбы с ними – отказ от ручного ввода. При работе с автоматизированной системой оператор устает значительно меньше, кроме того, степень его усталости почти не влияет на качество ввода, поскольку в ABBYY FormReader заложена возможность

автоматической проверки данных. Даже если оператор совершит какую-либо ошибку, система, обнаружив несоответствие со словарным словом (или иным эталоном), выдаст соответствующее предупреждение.

- **Ошибки распознавания.** В процессе распознавания система помечает некоторые символы как «неуверенно распознанные». Они передаются оператору для дополнительной проверки. Однако если программа, проводя распознавание, самонадеянно остановит свой выбор на неверной гипотезе, такой символ на ручную проверку не попадет, а сразу поступит в экспортный поток. Таким образом, информация окажется искаженной! Это самая серьезная проблема всех систем автоматизированного ввода. При разработке ABBYY FormReader 6.5 специалисты ABBYY уделили особое внимание борьбе с такими скрытыми ошибками. Как показывают тесты, вероятность, что такого рода ошибка останется незамеченной, удалось снизить до 0,5% при распознавании букв и до 0,1% – при распознавании меток.

**Подведем итог.** Одна из основных задач, которую решает система, – повышение качества вводимых данных; в этом нам помогают

- предварительная обработка изображений;
- проверки по типам данных;
- удобная система верификации данных;
- проверки формата данных;
- логический контроль данных;
- проверка сборки многостраничных документов (ABBYY FormReader 6.5 Enterprise Edition).

## Предварительная обработка изображений

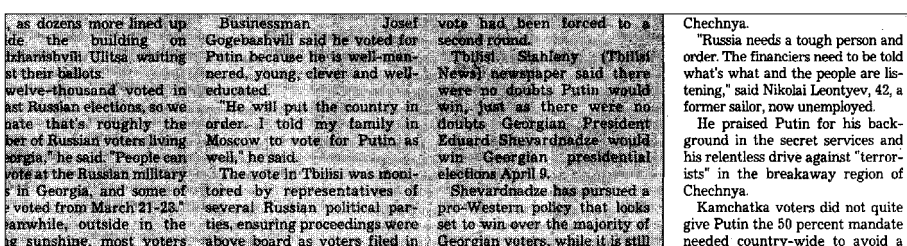
Нередко на изображениях форм присутствует «мусор» – точки разного размера, иногда листы сканируются под углом и изображение получается несколько перекошенным. Бывают ситуации, когда формы при сканировании развернуты на 90 градусов. Для системы распознавания крайне важно минимизировать воздействие такого рода факторов.

ABBYY FormReader 6.5 умеет делать следующее:

- очищать изображение от мелкого «мусора», причем имеется возможность задавать в интерфейсе размер «мусора», подлежащего удалению;
- исправлять перекошенные изображения с углом отклонения до 10 градусов;
- поворачивать страницы на угол, кратный 90 градусам;

- осуществлять инверсию – операцию преобразования негатива в позитив или наоборот.

Умеет программа отслеживать фоновое изображение (так называемую текстуру), состоящее из точек или произвольных линий, имеющих толщину гораздо меньше, чем в распознаваемых элементах. ABBYY FormReader обнаруживает и удаляет текстуру с изображения непосредственно перед анализом текста и распознаванием. Если система встречает множество отдельных небольших точек, то она удаляет их ещё на этапе предварительной обработки, а если – сетку из довольно длинных тонких линий, то их отделение и удаление производится уже при определении структуры документа.



Пример изображения, содержащего текстуру.

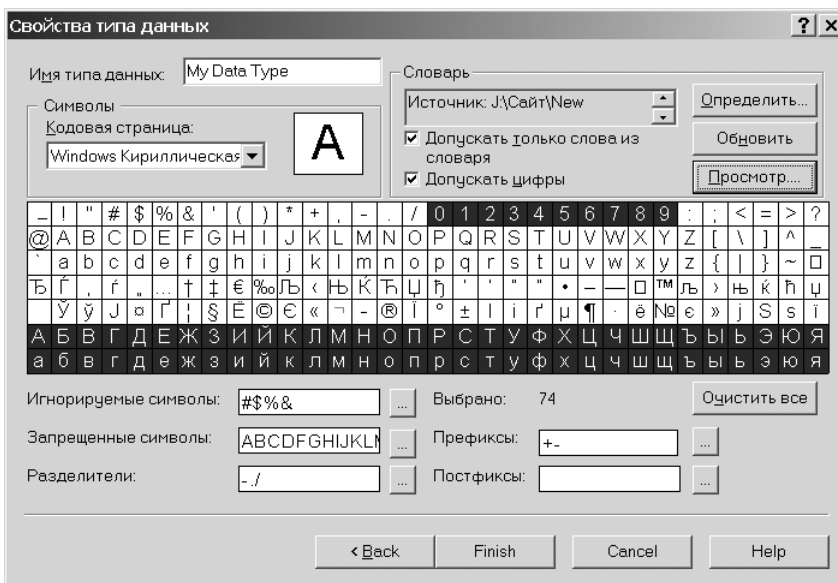
## Проверка по типам данных

Ещё до верификации, непосредственно в процессе распознавания, ABBYY FormReader 6.5 проводит **проверки по словарям и пользовательским базам данных**.

Допустим, в нашей форме есть поле «Любимый сорт сыра» и у нас имеется список названий всевозможных сортов сыра. В такой ситуации мы можем создать новый тип данных «Сорт сыра» и указать, что в данном поле могут встречаться только слова из имеющегося перечня сыров. Использование такого рода списков в программе ABBYY FormReader называется подключением словарей и помогает программе выбрать правильный вариант. Когда для поля в шаблоне указано соответствие

определённому словарю, программе распознавания легче ориентироваться при принятии решения.

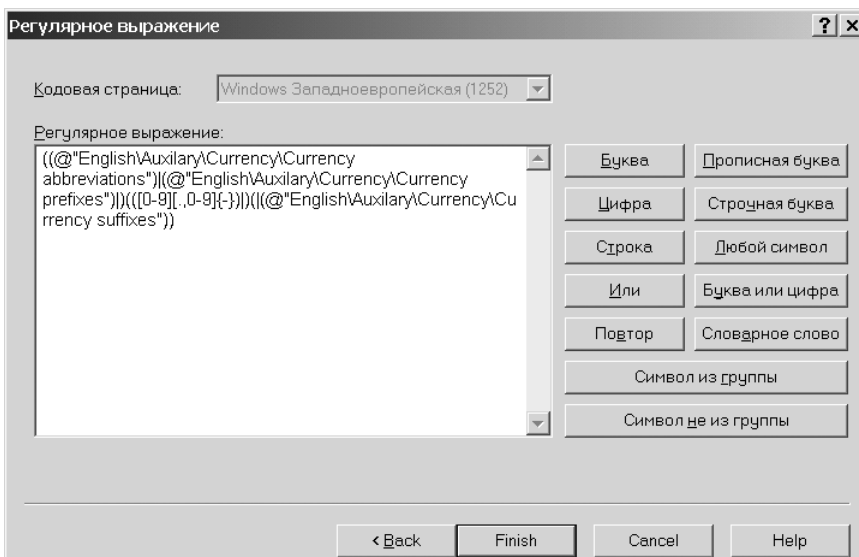
В комплекте ABBYY FormReader 6.5 поставляется набор стандартных типов данных, тематика которых охватывает все наиболее актуальные области. В частности, в набор входят словари русских имён, русских фамилий, названий российских городов и т.п. Аналогичные словари подготовлены и для многих других языков. Понятно, что заранее разработать словари на всё случаи жизни нереально. Зато пользователь ABBYY FormReader 6.5 может создавать собственные типы данных и подключать любые словари.



Создание пользовательского типа данных и подключение словаря в ABBYY FormReader 6.5 Desktop Edition

Наряду со словарными типами данных, широко применяется определение типа данных на основе регулярного выражения. Регулярное выражение определяет возможные комбинации

символов и их взаимное расположение. Например, регулярное выражение «к \* т» допускает все трёхбуквенные слова, в начале которых стоит «к», а в конце «т» – «кит», «кот» и т.п.

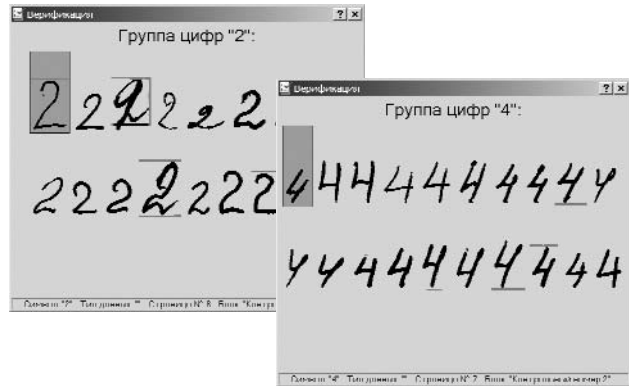


Подключение типа данных на основе регулярного выражения в ABBYY FormReader 6.5 Desktop Edition

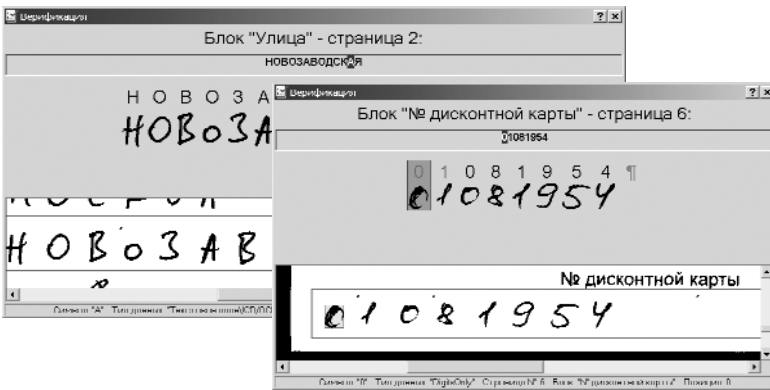
## Верификация

Поскольку точность распознавания форм произвольного вида всегда несколько ниже 100%, для повышения качества ввода в ABBYY FormReader 6.5 реализована верификация – проверка распознанных данных человеком. В системе реализовано три способа верификации.

1. **Групповая проверка.** Идеально подходит для одновременной проверки данных, заведомо принадлежащих к относительно небольшому множеству, например, цифр. При групповой проверке все неуверенно распознанные символы одного вида (скажем, все тройки) выводятся на экран перед оператором. В силу особенностей человеческого восприятия верификатору проще выделить один нетипичный символ из большого количества однотипных, чем искать тот же самый символ в тексте. Понятно, что скорость верификации за счёт применения групповой проверки существенно возрастает – ведь оператор может одним нажатием на кнопку Enter подтвердить сотни символов!



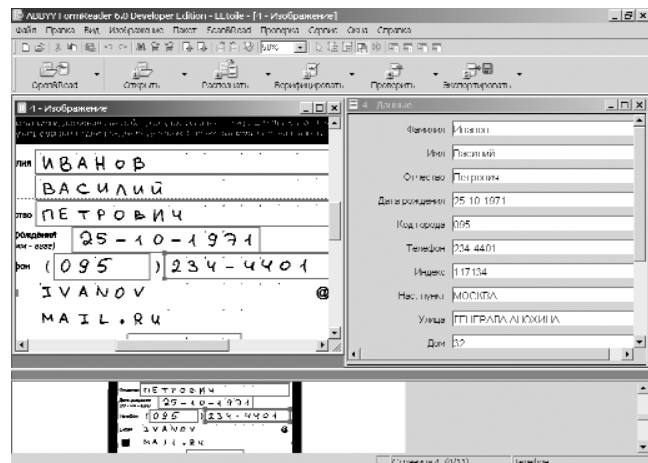
Проверка цифр при помощи групповой верификации.



2. **Контекстная проверка.** На экран одновременно выводятся две строки – фрагмент исходного изображения и результат распознавания. Таким образом оператор может сличать результаты распознавания с содержимым поля. Оператор подтверждает правильность распознавания (нажатием всего одной клавиши) либо исправляет неверно распознанные символы.

Контекстная проверка неуверенно распознанных символов.

3. **Проверка в форме.** Если те или иные контрольные проверки свидетельствуют о наличии серьёзных ошибок, форма, то такая помечается «флажком». Затем форма выводится на экран, чтобы верификатор поочерёдно осмотрел все поля «подозрительной» формы и внес необходимые изменения.



Проверка данных в форме.

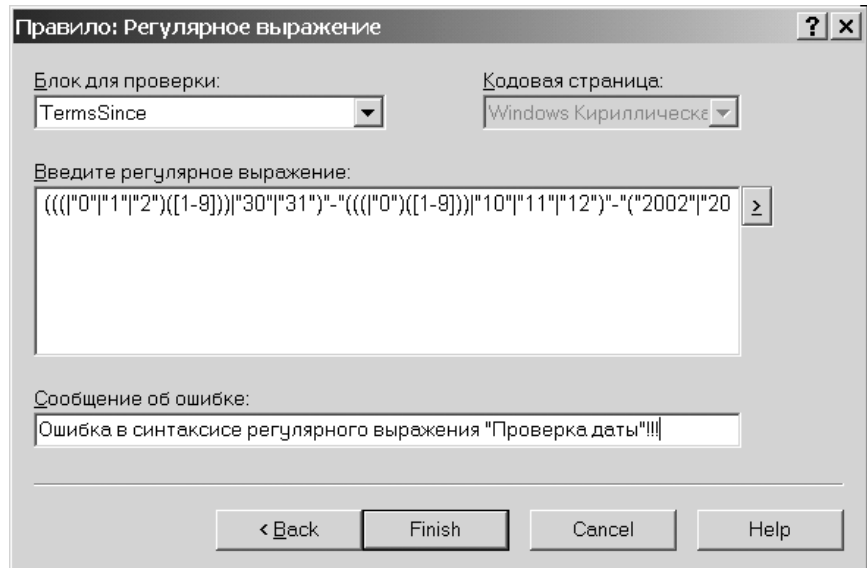
Все вышеописанные способы верификации реализованы в ABBYY FormReader 6.5 в рамках общей идеологии, суть которой – **минимизировать количество нажимаемых верификатором клавиш**. Именно этот параметр, а вовсе не количество неуверенно распознанных символов, серьёзнее всего влияет на скорость и качество верификации и, следовательно, на общее качество ввода. Ведь зачастую почти все символы, подсвеченные как «подозрительные», системой распознаны правильно. А чтобы подтвердить название переулка «2-ой Спасоаливковский» из более чем 20-ти подсвеченных, но правильно распознанных символов, требуется ОДИН раз нажать на клавишу Enter.

## Проверка формата данных

По окончании распознавания ABBYY FormReader 6.5 проверяет распознанные данные на соответствие указанным при создании шаблона форматам. Рассмотрим этот вид контроля на примере поля «Серийный номер». Предположим, известно, что номер должен состоять из сочетания SNFR, за которым должна следовать одна цифра, затем ещё две группы цифр:

**SNFRn-*nnn*-nn**

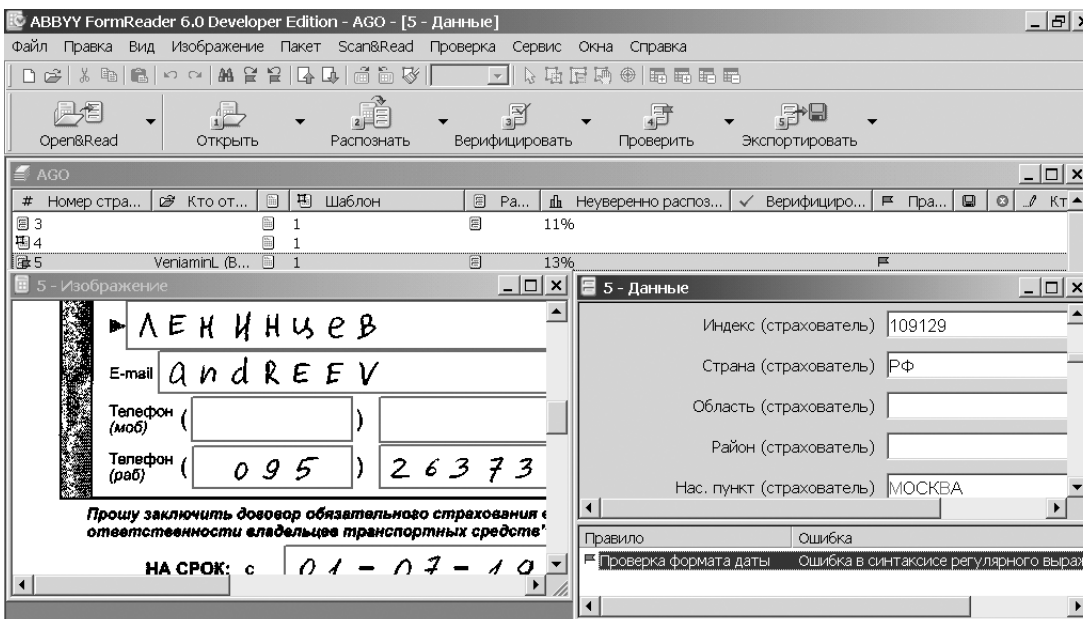
Разрабатывая шаблон, пользователь может определить специальное **правило проверки с помощью регулярного выражения**, имеющего следующий вид:



Определение правила проверки при помощи регулярного выражения.

Система, анализируя данные, полученные для поля «Серийный номер», **поставит флажок ошибки** те страницы, где количество цифр не совпадает с заданным, где вместо цифр оказались буквы и т.д. Это позволяет быстро и точно выявлять такие сложные для визуального различения ошибки, как русская

буква «О» вместо цифры «0», буква «З» вместо цифры «3» и т.п. В соответствующих случаях система выдает предупреждение или сигнал об ошибке. В последнем случае страница не может проходить дальнейшую обработку, пока оператор не внесет необходимые правки.



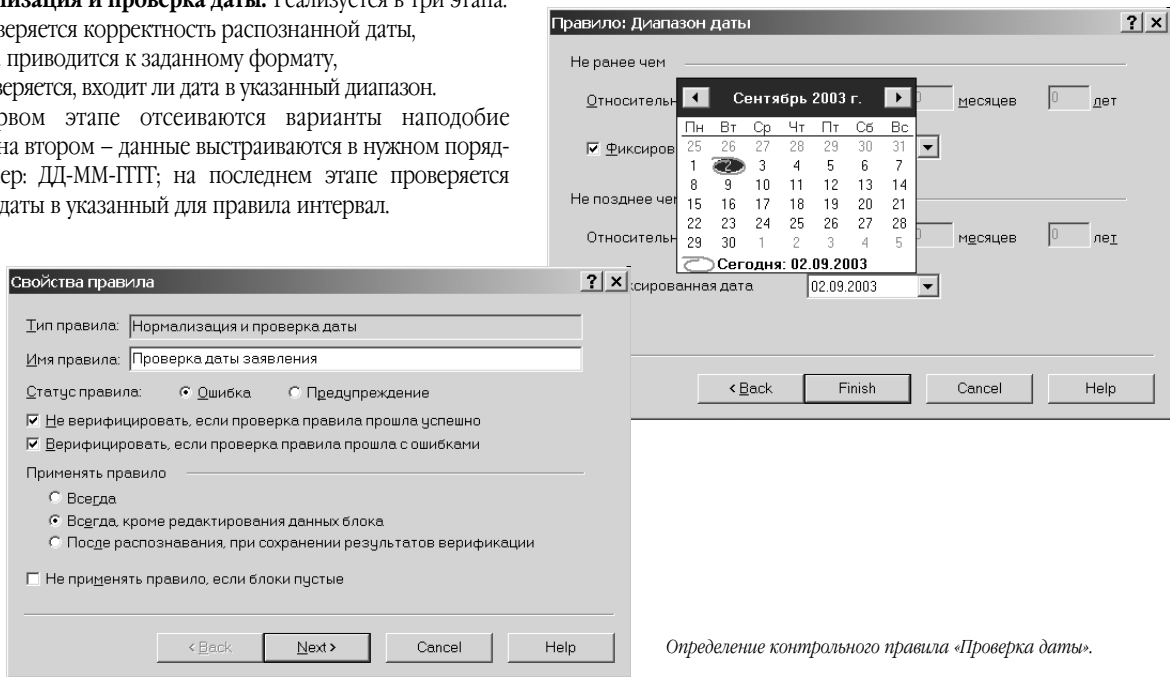
На странице была обнаружена ошибка – не выполняется контрольное правило «Проверка формата даты».

## Логический контроль

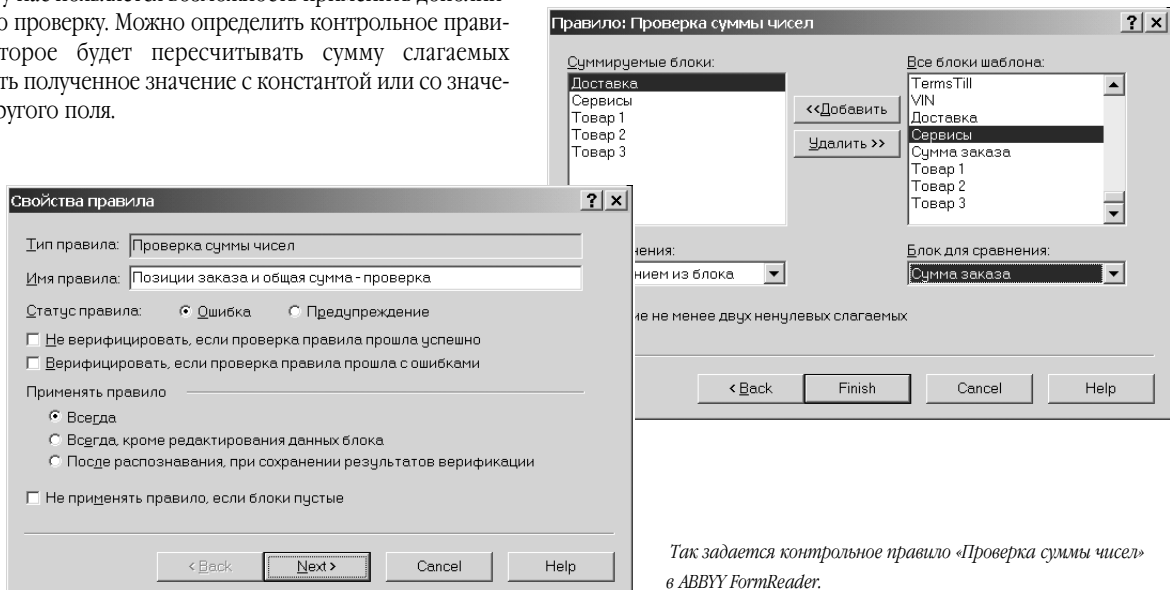
Данные в форме зачастую имеют какие-либо ограничения. Например, про них известно, что они должны попадать в определенный интервал. В этом случае после распознавания данные проверяются на выполнение наложенного условия и, если оно не выполняется, оператору выдается сообщение об ошибке. Рассмотрим некоторые правила.

- **Нормализация и проверка даты.** Реализуется в три этапа:
  - проверяется корректность распознанной даты,
  - дата приводится к заданному формату,
  - проверяется, входит ли дата в указанный диапазон.

На первом этапе отсеиваются варианты наподобие 32/45/199; на втором – данные выстраиваются в нужном порядке, например: ДД-ММ-ГГГГ; на последнем этапе проверяется вхождение даты в указанный для правила интервал.



- **Проверка суммы чисел.** Если на форме есть несколько числовых полей, сумма которых также присутствует на форме, у нас появляется возможность применить дополнительную проверку. Можно определить контрольное правило, которое будет пересчитывать сумму слагаемых и сверять полученное значение с константой или со значением другого поля.



- **Нормализация цены.** Строка «цена» автоматически приводит цену к заданному виду (например, 12,90 или 12,90). Система сообщает об ошибке, если распознанное значение невозможно преобразовать в нормализованный вид.
- **Проверка с условием.** На специальном языке, напоминающем языки программирования высокого уровня, пользователь может описать требуемое условие и определить действия, которые следует произвести в случае выполнения или невыполнения. Ниже приводится пример описания условия, согласно которому будет выдано сообщение об ошибке, если поля «страна» или «город» окажутся незаполненными.
- **Проверка соответствия числа цифрами числу прописью.** Сопоставляет значение, распознанное в блоке «число цифрами», со значением в блоке «число прописью». Данное правило может быть применено только к целым числам в русскоязычных документах.
- **Automation-проверка.** Правило проверки через OLE Automation дает возможность пользователю задать свои собственные, сколь угодно сложные правила проверки, при невыполнении которых будет выдаваться сообщение об ошибке.

```
if [City]. IsEmpty() then Error ([City],  
«Не указано название города»)  
else  
    if [Country]. IsEmpty() then Error( [Country],  
«Не указано название страны»)  
    else TRUE
```

Таким образом, система автоматизированного ввода форм в состоянии обеспечить разнообразный логический контроль за распознаваемыми данными: автоматически отыскать вкрапившиеся ошибки, указать на них оператору и не пропускать данные до тех пор, пока они не станут соответствовать заданным ограничениям.

## Обработка многостраничных форм

При обработке многостраничных форм случаются ситуации, когда страница из одной формы попадает в другую форму. Это серьезная проблема, ставящая под угрозу качество вводимых данных. Во избежание ошибок подобного рода на многостраничных формах предусматривают специальное поле – **уникальный идентификатор**. На всех страницах одной формы проставляется, естественно, один и тот же идентификатор. По этому полю система определяет принадлежность отдельных страниц к документу и производит так называемую сборку.

Для обработки многостраничных форм с меняющимся составом страниц разработана система потокового ввода ABBYY FormReader Enterprise Edition. Эта версия имеет ряд принципиальных отличий от Desktop Edition. В частности, в ней предусмотрен механизм описания форм, состоящих из нескольких страниц.

Применительно к работе ABBYY FormReader Enterprise Edition сборкой называют составление из данных, полученных с разных, возможно, разрозненных страниц, единой информа-

ционной структуры, ассоциируемой с конкретной многостраничной формой.

Примером идентификатора, успешно используемого при потоковом вводе большого количества многостраничных форм, может выступить ИНН (идентификационный номер налогоплательщика). В налоговых декларациях физических лиц ИНН указывается на каждой странице, что позволяет безошибочно собирать вместе страницы налоговой декларации.

Если по каким-то причинам правило сборки многостраничного документа выдает ошибку, документ подается на станцию коррекции – специальную рабочую станцию, где производится анализ ситуации. После чего оператор вручную переставляет страницы одного или нескольких документов, добавляет или заново сканирует отдельные страницы, запрашивает повторный ввод всего документа и т.д., а затем еще раз проверяет сборку, запуская правила проверки.

## Спокойная работа оператора – еще одна гарантия качества!

Самое главное, что даёт внедрение системы автоматизированного ввода – **возможность избавить сотрудников от тупой, монотонной работы** по ручной «набивке» данных. Нашему оператору нужно заниматься на 90% верификацией, т.е. в большинстве случаев нажимать на одну-две кнопки. При этом ему не приходится постоянно переключать внимание с бумаги на клавиатуру и экран и обратно; он не должен каждый раз вспоминать, в какую же графу базы данных следует внести ту или иную цифру. Все нужные логические связи продуманы,

проверены и отлажены ещё на этапе внедрения системы. В результате, дело продвигается быстрее и спокойнее.

Но самый важный эффект от внедрения всё же заключается не в этом. Когда люди перестают сильно уставать, ежедневно портить глаза, нервничать из-за срыва сроков, так или иначе улучшается настроение каждого работника, общая атмосфера в коллективе. А подобные улучшения стоят очень дорого. И к тому же – не покупаются.

## Как правильно организовать автоматизированный ввод документов

Итак, если принять во внимание качество получаемых данных, удобство работы оператора и скорость обработки документов, автоматизированный ввод данных имеет несомненные преимущества перед ручным. Экономически оправданной автоматизация ввода становится при обработке от 100 и более документов в день.

Чтобы автоматизировать ввод даже при небольших объемах, потребуются некоторые изменения в организации работы операторов. Когда же объем ввода достигает нескольких тысяч форм ежедневно, автоматизация ввода становится задачей производственного масштаба и требует ощутимых организационных усилий.

### Подходы к организации потокового ввода данных

Выделяют два основных подхода к организации потокового ввода данных: обработка форм по мере их поступления и обработка форм по мере накопления. Соответственно, система автоматизированного ввода внедряется либо во фронт-офисе (секретариате, клиентском отделе), либо в бэк-офисе (вычислительном центре, внутренних отделах).

#### Ввод данных во фронт-офисе

Характерным примером первого подхода может служить система ввода форм, внедрённая на складе торговой организации. Представитель организации-заказчика, обращаясь на склад, заполняет форму, в которой указывает, какой именно товар и в каких количествах он желает получить. Понятно, что такая форма должна быть обработана сразу, как только она поступит к сотрудникам склада; на основе этой формы выписывается счет, который клиент оплачивает, пока кладовщики и грузчики готовят заказ. Поэтому система автоматизированной обработки устанавливается непосредственно по месту приёма форм, на складе. Исходя из очевидных требований к функционированию такой системы, можем назвать ряд её особенностей.

1. Скорость сканирования в данном случае не слишком важна – лишние две-три минуты на обработку каждой формы не замедлят общий процесс, поскольку погрузо-разгрузочные работы всё равно занимают намного больше времени. Следовательно, в составе системы может работать любой сканер, даже дешёвый планшетный аппарат из тех, что используются в офисах. Все современные сканеры способны оцифровать лист формата А4 за 30–40 секунд, что вполне приемлемо в описанной ситуации.

2. Полный цикл операций ввода проводится на одном рабочем месте – непосредственно там, где клиент заполняет и сдаёт форму. Более того, там же осуществляются все вспомогательные операции, не связанные прямо с вводом данных: в нашем примере это оформление счёта, утилизация бумажной формы или помещение её в архив и т.п.

Очевидно, что в подобных случаях для организации ввода форм требуется система класса ABBYY FormReader Desktop Edition.



Ввод форм по мере их поступления.

## Ввод данных в бэк-офисе

Проиллюстрировать второй подход – «обработка по мере накопления» – легко на примере налоговых деклараций. Как известно, Государственная налоговая служба России организовала сбор и обработку налоговых деклараций частных лиц следующим образом: определённое время (несколько месяцев) ведётся приём деклараций и только потом они централизованно обрабатываются. В результате накапливаются большие объёмы документов. Для их автоматизированного ввода необходима система промышленного уровня. Аналогично предыдущему случаю можем сформулировать основные особенности такой системы.

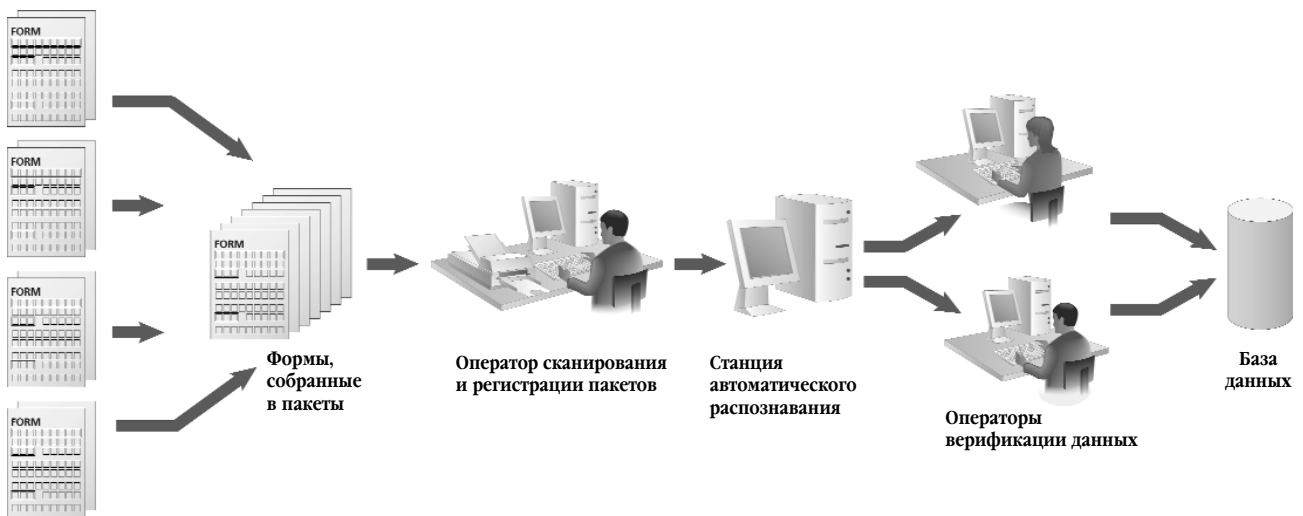
1. Необходимо использование промышленных, высокопроизводительных сканеров.
2. Должна быть организована распределённая система обработки. Каждый из операторов ввода должен иметь чёткую специализацию и на всём протяжении процесса обслуживать одну рабочую станцию: сканирования, распознавания, верификации или экспорта.
3. Требуется значительно более строгий, чем в предыдущем примере, контроль качества. Причины очевидны: если состав заказа всегда может быть легко скорректирован, то проделать нечто аналогичное с информацией, содержащейся в налоговых декларациях миллионов граждан, практически невозможно.

4. Весьма желательно организовать двух- или трёхсменную работу вычислительного центра. Это, помимо прочего, позволит максимально использовать возможности высокопроизводительных сканеров, рассчитанных на круглосуточную работу.

Несложно убедиться, что в данном случае лучше всего подходит система класса ABBYY FormReader Enterprise Edition.

Сравнительный анализ этих примеров приводит к заключению о том, что для каждого конкретного проекта может быть подобрана оптимальная конфигурация системы автоматизированного ввода. Однако решение о выборе конфигурации должно приниматься с учётом всех особенностей проекта, а также возможностей системы. Поэтому обычно для разработки плана внедрения той или иной системы необходимо привлечь специалистов как компании-разработчика, так и представители компании-заказчика, располагающих полной информацией об особенностях работы данного предприятия.

Рассмотрим основные принципы построения больших систем автоматизированного ввода форм, позволяющие добиться высокой эффективности при сохранении других параметров ввода (качество данных, скорость обработки и т.д.). Все принципы сформулированы на основе опыта успешных масштабных проектов по вводу форм и положены в основу программных продуктов компании ABBYY.



Входящие формы

*Сбор, накопление форм и их последующий потоковый ввод.*

## Основные принципы потокового ввода

### Пакетная обработка данных

Смысл этого принципа состоит в том, что однотипные формы в рамках системы объединяются в так называемые пакеты. Иными словами, на программном уровне однотипные формы рассматриваются как содержимое некоторого обособленного контейнера. Каждый такой пакет имеет уникальный идентификатор. Преимущества очевидны: подобное решение позволяет структурировать поток вводимых данных; каждый пакет может иметь свои программные настройки; облегчается администрирование, маршрутизация и хранение данных.

### Распределение функций операторов

Мощные системы ввода данных обычно функционируют по принципу конвейера. Смысл такого построения понятен: специализация повышает производительность труда, а также позволяет практически неограниченно масштабировать систему. Например, всегда можно увеличить количество мест операторов сканирования, никаким образом не вмешиваясь в работу операторов распознавания, верификаторов и т. д.

### Масштабируемость системы

Как показано выше, благодаря распределению функций между операторами, система оказывается состоящей из узкоспециализированных модулей. При этом количество модулей каждого вида определяется только особенностями конкретной ситуации и может быть при необходимости легко изменено. Например, на один модуль распознавания (мощный 2-процессорный сервер) изначально приходилось 8 модулей верификации; через какое-то время выяснилось, что верификация стала «узким местом» системы; решили добавить ещё 4 таких модуля, после чего информационный «затор» был устранён. Понятно, что все эти особенности делают систему более гибкой и управляемой, существенно удешевляют комплекс в целом.

### Очередность заданий

Важным для потокового ввода является понятие маршрута движения пакета, так называемой магистрали. Пакеты движутся по системе не произвольным образом, а в соответствии с заданной схемой.

В каждый момент времени пакет имеет определённый статус, указывающий, на какой стадии обработки он находится. В сложных системах статус, скажем, «на верификацию» могут одновременно иметь 5–10 пакетов, составляющих так называемую очередь. Как только одна из станций верификации освобождается, следующий пакет из очереди автоматически переправляется на обработку. Основное преимущество системы, использующей очереди заданий, – равномерное распределение нагрузки по всем ресурсам (операторам) системы. Как только, например, оператор верификации освободился от задания и сообщил о своей готовности продолжить работу, на его рабочее место доставляется следующий пакет форм.

### Сохранение магистрали ввода

До тех пор пока обработка документа идёт без серьёзных затруднений, он обрабатывается в общем потоке. Если же возникают какие-либо проблемы, например из-за сбоя при сканировании, то документ должен быть немедленно исключён из магистрали, чтобы не замедлять прохождение остальных пакетов. Как правило, «проблемные» пакеты передаются на ручную обработку – оператор, установив причину возникновения сбоя, выбирает способ решения проблемы. В нашем примере документ будет направлен на повторное сканирование. Заметим, что обработка остальных пакетов всё это время продолжается в прежнем, высоком темпе.

## Проект по промышленному вводу форм

Особого рассмотрения заслуживает проект по вводу форм в промышленном масштабе. Для автоматизации такого проекта требуется специальное программное и аппаратное обеспечение, обучение операторов и организация их работы.

### Программное решение для ввода форм

Практика показывает, что если необходимо в день обрабатывать более 3000 документов и привлекать для этого более трех сотрудников, то максимальная эффективность достигается при распределении этапов обработки. Тогда каждый сотрудник будет иметь возможность сосредоточиться на определенной операции и выполнять ее быстро и качественно. Именно такой принцип работы заложен в программном продукте ABBYY FormReader Enterprise Edition.

В сочетании с уже упомянутыми возможностями связывать многостраничные документы, применять разнообразные правила, система подобного ранга позволяет организовать процесс практически без ограничений по объему ввода и для форм любой сложности

### Промышленный сканер

Обязательно использовать высокопроизводительные сканеры. Иногда теоретически рассматривают альтернативное решение, заключающееся в распределении входного потока форм по большому количеству (десяткам и сотням) операторов сканирования, работающих на дешёвых аппаратах с невысоким быстродействием. Однако практика показывает, что подобный подход нереализуем: сканеры, не рассчитанные на потоковый ввод тысяч документов в день, имеют намного меньшую нагрузку на отказ, нежели промышленные аппараты. Соответственно, через короткое время организаторы столкнутся с массовой поломкой сканеров и дополнительными затратами на их ремонт или переоборудование рабочих мест.

### Аппаратное обеспечение

- Для станций сканирования годятся практически любые компьютеры, нужно лишь правильно оценить требуемый для сохранения отсканированных изображений объём жёсткого диска.
- Станции распознавания должны иметь большую вычислительную мощность, для чего понадобится процессор с высокой тактовой частотой и достаточный объём оперативной памяти. В роли станций распознавания часто используют многопроцессорные серверы. Большинство систем ввода, в частности ABBYY FormReader, поддерживают многопроцессорную обработку.

- Для верификаторов, нужны качественные мониторы; в противном случае у операторов будут быстро уставать глаза, что неизбежно понизит общее качество ввода.
- Компьютер, на котором будет работать станция экспорта, должен иметь достаточный ресурс оперативной памяти, поскольку на нем в фоновом режиме будут экспортироваться данные.
- Сама по себе локальная сеть в отделе автоматизированного ввода форм обязательно должна иметь высокую пропускную способность – внутренний трафик, как правило, достаточно велик. Для сравнения скажем, что объём одного пакета может достигать десятков Мбайт.
- Так как для хранения настроек комплекса и вводимых данных ABBYY FormReader Enterprise Edition использует внутреннюю базу данных, требуется достаточно мощный компьютер, используемый как сервер базы данных.

### Организация экспорта данных

Обычный для настольных систем экспорт распознанных данных в файл указанного формата для систем промышленного ввода непригоден. При создании любой более-менее масштабной системы обычно пишут специальный модуль экспорта. Последний позволяет организовать поточную передачу информации во внешнюю систему обработки и хранения. В качестве альтернативы иногда рассматривают экспорт в файл формата XML, который поддерживается во всех продуктах линейки ABBYY FormReader 6.5, с последующим применением специально разработанного анализатора XML-файлов.

### Обучение персонала

Прежде чем приступить к работе с подобной системой, надо провести инструктаж персонала. Хотя интерфейс ABBYY FormReader Enterprise Edition разработан таким образом, чтобы оператор мог работать не задумываясь и не совершая ошибок, краткий курс обучения сотрудникам не повредит. Поэтому, участвуя в подобных проектах, специалисты компании ABBYY всегда оказывают содействие в создании инструкций для персонала, а также проводят обучение операторов всех специализаций. Особенно эффективным оказывается обучение непосредственно на рабочих местах – ведь это помогает операторам легко и быстро войти в работу!

# Решение нетривиальных задач с помощью технологий ABBYY

В некоторых случаях применение продуктов ABBYY кажется, на первый взгляд, неэффективным. Рассмотрим несколько таких ситуаций.

## Если система не поддерживает распознавание языка документа

Представьте себе, что язык, на котором заполняются формы, не поддерживается выбранной системой ввода. Например, корейский или тайский языки, которых нет в ABBYY FormReader.

Также известно, что написанный слитно, не разделённый на буквы текст система распознавания, анализирующая отдельные буквы, обработать не сможет. Обе ситуации схожи: система, в силу ограниченности своих возможностей, не распознает буквы в поле, но оператор может их прочитать.

В таких случаях рекомендуется применять следующее.

1. При разработке формы **как можно меньше использовать текстовые поля, заменяя их метками или группами меток**. Пояснительные надписи при метках могут быть выполнены на любом языке – эти надписи не подвергаются распознаванию.
2. Выделить **цифровые поля** и поля, содержащие **штрих-коды**, – для этих полей, скорее всего, удастся реализовать автоматическое распознавание.
3. Привлечь механизм **Key From Image (KFI)**, или «поля для ручного ввода», как они названы в продуктах линейки ABBYY FormReader 6.5. Так принято называть ввод данных

оператором, без автоматического распознавания. На экран при этом выводится извлечённое из отсканированного документа изображение текстового поля, и оператор может, читая написанный текст, перепечатывая соответствующую информацию. Заметим, что при этом сохраняются почти все основные преимущества автоматизированного ввода данных. Действительно:

- во-первых, все смысловые связи между информацией и полями итоговой базы данных продуманы заранее и жёстко зафиксированы в шаблоне формы, поэтому оператору не нужно размышлять на тему «что куда писать»;
- во-вторых, ему нет необходимости распределять внимание между бумажным документом, клавиатурой и экраном;
- в-третьих, сохраняется возможность задействовать любые, сколь угодно сложные и разветвлённые алгоритмы автоматической проверки данных – для проверочных правил не имеет значения, поступила информация после распознавания или с клавиатуры.

Таким образом, ввод данных при помощи «полей для ручного ввода», оказывается всё равно **значительно более быстрым и точным**, чем ввод ручным способом.

정 약 사 항				우체국보험청약서				(우체국 제출용)			
보험종류	어깨동무2종 코드( 5123 )			구 분	보험가입금액	보 험 료		증서(청약)번호			
보험기간	( )년, ( 80 )세, ( )종신			주 계 약	1,000 만원	19,700 원		04707079110			
납입기간	( 5 )년, ( )세, ( )일시납			특 약	만원	원		계 약 국	영등포우체국		
수금방법	①방문      ②창구			약	특약	만원 원		모 집 자	오창경 코드 010108		
	2      ③우체국이체      ④은행이체				특약	만원 원		모집구분	①창구      ②개인		
	연금계시연령 ( )세				합 계 보 험 료	19,700 원		세금우대	2      ①신청      ②미신청		
부활계약	부활국	(국기호)	수금자	(코드)	부활기간	원		부활보험료	원	부활이자	원
단체계약	단체번호		신규여부	①신규      ②추가	※ 일반단체의 경우 별첨 "일반단체 피보험자 청약명세서"에 기재하여 주십시오.						
구 분	성 명	주민등록번호		관계	주 소						
	(단체명)	(사업자등록번호)									

Даже если ABBYY FormReader не поддерживает языка документа, использование программы может сделать ввод данных быстрее.  
Пример документа на корейском языке.

## Удалённое сканирование и обработка факсимильных документов

Входное изображение необязательно должно поступать со сканера. Если по каким-либо обстоятельствам нет возможности внедрить систему автоматизированного ввода в том месте, где заполняются и собираются формы, эта особенность ABBYY FormReader может оказаться весьма полезной.

Рассмотрим простой пример: в нескольких городах области проводятся опросы населения. Отдел автоматизированной обработки данных организован при областной администрации и находится в областном центре. Понятно, что транспортировки тысяч собранных форм из разных городов лучше избежать. Если предположить, что организаторы опросов используют систему ABBYY FormReader, проблема решается элементарно. На местах, в пунктах сбора форм, документы сканируют и сохраняют (например, в формате чёрно-белый TIFF – каждый такой файл будет иметь небольшой объём, 10-100 кбайт), затем высылают по электронной почте в центр обработки. В центре, где расположена собственно система

автоматизированного ввода, файлы сохраняют и передают на станцию распознавания. ABBYY FormReader 6.5 способен принимать такие файлы и автоматически формировать из них пакеты для последующей обработки.

Аналогично, вместо электронной почты в такой ситуации может использоваться факсимильная связь. Несмотря на то что передача данных по факсу может быть сопряжена с линейными искажениями – растяжением, сжатием и т.п., ABBYY FormReader способен корректировать и правильно обрабатывать такие изображения. Единственное требование, налагаемое на формы, которые планируется передавать по факсу, – на них должны быть предусмотрены реперные блоки в виде чёрных квадратов, расположенных по углам формы. Налагая шаблон, система ориентируется на то, насколько изменилось взаимное расположение черных квадратов. В зависимости от этого корректируется алгоритм поиска полей, содержащих распознаваемую информацию.

## Распределённая верификация

Высокой квалификации для работы верификатора не требуется. Нужен только компьютер, причём не самый мощный, и внимательность. Вполне естественно, что многие организаторы для снижения себестоимости проекта потокового ввода форм привлекают в качестве верификаторов людей, работающих на дому. Единственным специальным требованием, предъявляемым к сотрудникам, в этом случае является наличие у них выхода в интернет.

В тех случаях, когда такой подход становится оправданным, компания ABBYY рекомендует применять для удалённой верификации системы терминального доступа. Система терминального доступа состоит из терминального сервера (ТС), который расположен на территории вычислительного центра, и терминального клиента (ТК), в роли которого выступает

компьютер домашнего работника. На сервере запускается программа, управляющая потоком направленных на верификацию данных; при этом клиентская машина не только «видит» происходящее в окне этой программы, но и позволяет оператору работать так, словно он находится за консолью сервера. Программа ТК перехватывает действия оператора, сообщает о них на ТС, серверная программа выполняет соответствующие операции верификации. Затем результаты выполненных действий транслируются обратно на ТК.

Продукты линейки ABBYY FormReader 6.5 протестированы для использования с системой терминального доступа Microsoft Terminal Services.

## Ввод «гибких форм»

Как уже упоминалось, существуют два основных типа форм: структурированные и гибкие. К структурированным относятся формы, расположение и размер полей которых фиксированы. Все остальные формы считаются гибкими.

Если структурированные формы удобнее и быстрее всего вводить методом наложения предварительно разработанного шаблона, то с гибкими формами при аналогичном подходе возникают известные затруднения. Специалистами компании ABBYY создан альтернативный метод обработки гибких форм, позволяющий решать задачу потокового ввода с качеством, не уступающим качеству при обработке структурированных форм. Суть метода заключается в логическом исследовании структуры документа; технология, реализующая данный метод на практике, получила название FlexiCapture.

Система, проводя анализ, определяет местонахождение и вид полей по признакам, заранее описанным пользователем. Допустим, требуется найти на форме и распознать содержимое поля «ИНН». В шаблоне указано, что признаком искомого поля являются буквы «ИНН», справа от которых находится известное количество знаменит. Тогда система обнаружит указанную комбинацию букв и без проблем справится с вводом соответствующей информации. Такой подход, в основе которого лежит исследование геометрически неопределенной структуры документа, включая взаимное расположение отдельных его элементов, принято называть Intelligent Field Recognition (IFR).

Специалистами компании ABBYY разработан метаязык для описания структуры и правил анализа документов произвольной формы. Возможности языка достаточно широки, чтобы охватить подавляющее большинство форм, используемых в настоящее время. Понятно, что это куда более мощный инструмент, чем традиционный шаблон для структурированных форм. Благодаря технологии FlexiCapture более 500 банков на территории России вводят поступающие платежные документы автоматизированно, при помощи программных решений от ABBYY.

Специализированный программный продукт ABBYY FlexiCapture Studio позволяет формировать описание гибких шаблонов, не прибегая к программированию. Для работы с FlexiCapture Studio достаточно обладать навыками работы с компьютером на уровне опытного пользователя.

Разработчик шаблона «обучает» программу FlexiCapture Studio искать нужные поля. Для этого в терминах FlexiCapture Studio создается описание расположения каждого поля формы через задание параметров его окружения: стационарного текста, рисунков, разделителей, белых пятен и т.д. Опираясь на описание, программа находит на форме все объекты такого рода и выбирает вариант, в наибольшей степени совпадающий с описанием. Если поля найдены правильно, описание тестируется на большом количестве форм, уточняется и переносится в программу ABBYY FormReader как шаблон для обработки бумажных документов.

TRANSCONTINENTAL TITLE CO 2605 ENTERPRISE RD. E. STE #200 CLEARWATER, FL 33759 1-800-789-2240		Instrument      Book Page 200100032776 OR 1369 1627	
20013540	MORTGAGE		
THIS MORTGAGE is made this 3rd day of JULY, 2001 between the Mortgagor, JEFF GASKINS AND MELODY GASKINS, HUSBAND AND WIFE HUSBAND AND WIFE			
and the Mortgagee, MILLENNIUM BANK, N.A., NATIONAL BANK a corporation organized and existing under the laws of VIRGINIA whose address is 1601 WASHINGTON PLAZA, RESTON, VIRGINIA 20190			
WHEREAS, Borrower is indebted to Lender in the principal sum of U.S. \$ 90,000.00 which indebtedness is evidenced by Borrower's note dated JULY 3, 2001 and extensions and renewals thereof (herein "Note"), providing for monthly installments of principal and interest, with balance of the indebtedness, if not sooner paid, due and payable on AUGUST 1, 2016			

## Ввод данных с немашиночитаемых форм

Специфика этой задачи заключается в том, что далеко не всегда требуется распознавать всё, что можно найти на изображении. Чаще всего при оцифровке архива немашиночитаемых форм оказывается достаточно получить электронное изображение страницы, выделить на нём и распознать некоторое количество ключевых полей. На основании этих ключевых полей для страницы строится уникальный составной индекс. Само изображение преобразуется в формат, удобный для хранения, и передается в архив. Если понадобится найти эту страницу, поиск будет значительно облегчен благодаря выделенному индексу.

Документы в архиве, скорее всего, не являются машиночитаемыми формами, поскольку

- расположение полей не определено,
- реперные блоки (квадраты, уголки и пр.) не предусмотрены,
- в полях могут находиться слитно написанные слова,
- поля могут перекрываться штампами и надписями.

Всё это означает, что общий подход, заключающийся в подготовке шаблона с заданной разметкой, неприменим.

### Что делать в такой ситуации?

1. Везде, где это возможно, применять стандартный подход с распознаванием данных **в полях с фиксированным расположением**. Для этого необходимо разработать шаблон документа. В качестве реперных блоков можно использовать таблицы или надписи, присутствующие на всех экземплярах. Обычно удается найти хотя бы десяток таких полей, и этого достаточно для уверенного наложения шаблона. В качестве идентификатора документа можно использовать, например, его название: «форма 4-ПД», «выписка о состоянии счета» и т.д. На накладной транспортной компании или на счете можно распознать штрих-код, заранее впечатанные данные о стране отправителя, телефон и почтовый код адресата.
2. Если поле «плавает», для определения его местоположения можно применить **технология FlexiCapture**. Этот подход позволяет найти любое поле, если имеется дополнительная

The image shows a UPS shipping form with the following visible data:

- Sender:** O. Feoktistova, 8 Butyrskaya str., Moscow, RUSSIA, 127015. Phone: 7 (095) 234 4400.
- Recipient:** OOO CE "ABECT", ul. Kuzbasa 89, Moscow, RUSSIA, 12078. Phone: 3472-231500.
- Service Level:** Express
- Weight:** 0.1 kg
- Insurance Value:** \$20.00
- Tracking Number:** M082 917 304 2

Документ, часть полей которого может быть распознана автоматически.

информация о его окружении, формате данных, поясняющих надписях и т.д. Единственная сложность – разработка «гибких» шаблонов является относительно дорогой услугой, которая может быть выполнена только подготовленным специалистом.

3. Иногда хорошие результаты достигаются при помощи **подхода Key From Image – «поля для ручного ввода»**. Его применение описано на стр. 26 для случаев, когда какие-либо языки не поддерживаются. Программа помогает найти положение поля на форме, «вырезает» на изображении участок, но данные не распознает. Оператор вводит их вручную. Таким образом можно ввести любую информацию с документов, не готовившихся специально для автоматизированного ввода. И это будет проще, чем при обычном ручном вводе.

The screenshot shows the ABBYY FormReader 6.0 Developer Edition interface. The main window displays the scanned form with various fields highlighted. A secondary window titled 'Данные' (Data) is open, showing the following extracted information:

- Tracking Number: M0829173042
- Shipper UPS Account No: 4a4334
- Company Name and Address: ABBYY SOFTWARE HOUSE, 8 BUTYRSKAYA Str.
- Shipper Postal Code: 127015
- City: Moscow
- Country: RUSSIA
- Phone: 3472-231500
- Postal Code: 454078
- Filing Date: 12/08/03

Автоматизированный ввод отдельных полей немашиночитаемой формы

## Заклучение

Мы рассмотрели важнейшие аспекты автоматизированного ввода данных с форм. Начав с описания основных понятий из этой области, мы перешли к пошаговому рассмотрению процесса, акцентируя внимание на наиболее важных, значимых его параметрах и особенностях. Мы затронули преимущества автоматизированного подхода перед ручным вводом, увидели, как можно влиять на качество вводимых данных, постарались осветить вопросы оптимальной организации ввода данных в различных практических ситуациях, привели советы по всем основным этапам подготовки бланка формы.

При подготовке материала мы опирались, прежде всего, на опыт применения технологий компании АВВУУ, однако при этом большинство советов и рекомендаций не имеют строгой привязки к какой-то конкретной линейке продуктов. Мы старались собрать в одном материале полезную информацию о такой области документооборота, как ввод форм, чтобы заинтересовать и профессионала и человека, впервые услышавшего об этом. Надеемся, что нам удалось добиться поставленной цели и приведенная информация окажется полезной всем заинтересованным в разработке, внедрении и эксплуатации подобных проектов.

## Контакты

### **ABBYY Software House (Москва)**

Тел: +7 095 783 3700  
Факс: +7 095 783 2663  
Россия, 129301, Москва, а/я 54  
formreader@abbyu.ru

### **ABBYY Europe GmbH**

Тел: +49-89-511159-0  
Факс: +49-89-511159-59  
Anglerstrasse 6, Munich,  
Germany, 80339  
sales@abbyeu.com

### **ABBYY Ukraine**

Тел: +380 44 490 9999  
Факс: +380 44 490 9461  
Украина, Киев, 02002, а/я 23  
sales@abbyu.ua

### **ABBYY USA**

Тел: +1 510 226 6717  
Факс: +1 510 226 6069  
47221 Fremont Boulevard,  
Fremont, California 94538, USA  
sales@abbyusa.com



© 2005 ABBYY Software Ltd. Все права защищены.  
ABBYY, FINEREADER, Scan&Read, ABBYY FineReader, FormReader, ABBYY FormReader – товарные знаки и зарегистрированные товарные знаки ABBYY Software Ltd.  
Adobe, Adobe Logo, Adobe PDF и Adobe Acrobat являются товарными знаками компании Adobe Systems Incorporated. Microsoft, Outlook, PowerPoint, Windows, Windows NT являются зарегистрированными товарными знаками или товарными знаками компании Microsoft Corporation в Соединенных Штатах Америки и/или других странах. Остальные товарные знаки являются товарными знаками или зарегистрированными товарными знаками своих законных владельцев.  
Part # 1591r  
ABBYY Software. 129301, Москва, а/я 54, тел.: (095) 783-3700, факс: (095) 783-2663, formreader@abbyy.ru, www.abbyy.ru, www.formreader.ru