Ольга Морозова,
руководитель направления корпоративных проектов ABBYY, Россия

Опыт внедрения потокового ввода документов

Переход на электронный документооборот подразумевает ввод документов в базу, то есть, их перевод в электронный вид. Обычно эта процедура производится вручную, причем не только в небольших, но и в достаточно крупных страховых компаниях. Компания ABBYY разработала и уже опробовала на практике инструментарий автоматизации ввода большого количества бумажных документов.

Автоматизация требует систематизации

Общеизвестно, что использование ручного труда при вводе данных из документов не обеспечивает уверенности в правильности введенных данных и достаточной скорости ввода. При этом ручной ввод информации с бумажных носителей в информационную базу данных компании чреват высокими трудозатратами, а не выявленные ошибки фактически становятся «слепыми» и способны дать непредсказуемые результаты. Кроме того, скорость

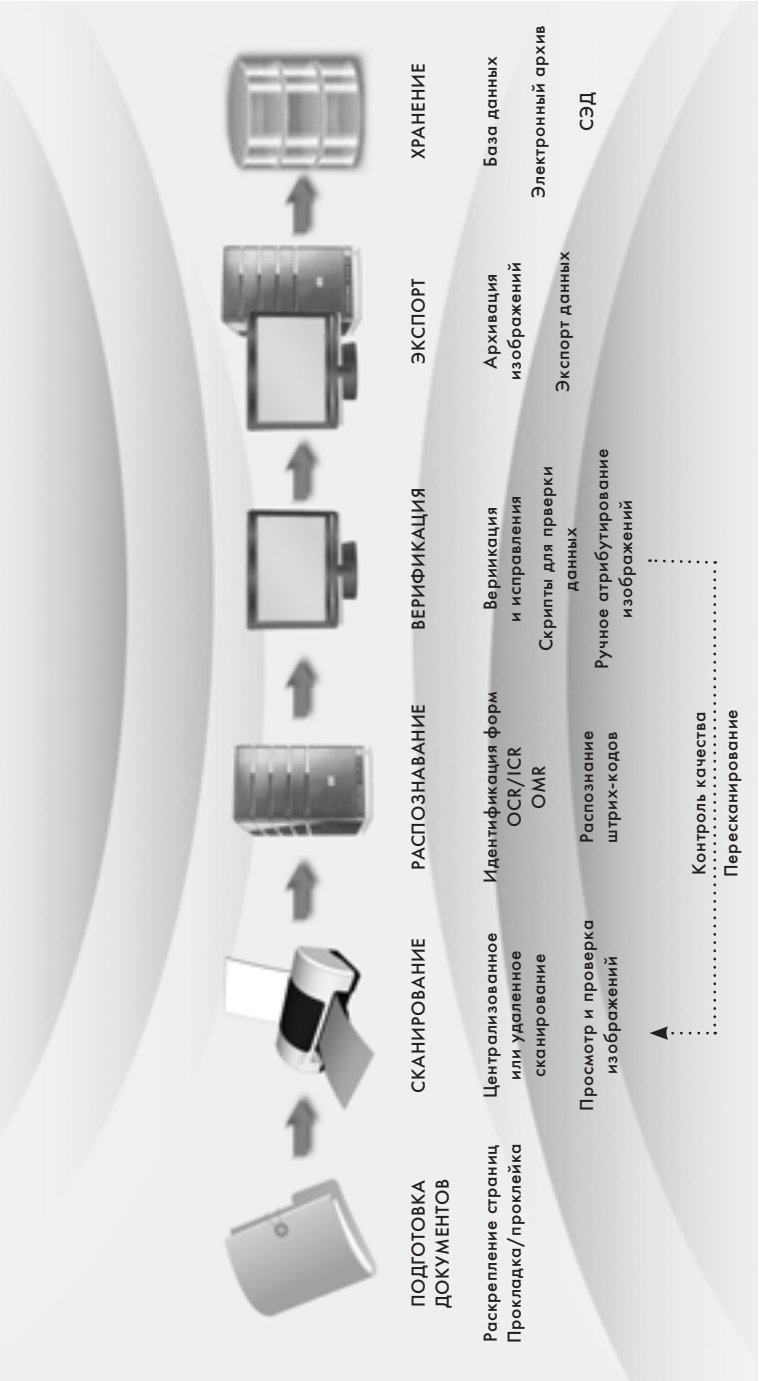

Рис 1. Схема организации потокового ввода документов

ручного ввода даже при использовании десятипальцевого метода набора данных нельзя назвать удовлетворительной. Однако несмотря на перечисленные недостатки далеко не все страховые компании – даже из первой десятки – находят возможность автоматизировать этот процесс. Между тем, переход на потоковый ввод данных позволяет не только упростить работу, повысить эффективность ввода данных и обеспечить сохранность копий оригинальных документов в системах электронного документооборота, но и способен выявлять совпадающие данные, гарантировать непротиворечивость вводимой информации, значительно снижать уровень ошибок и упрощать контроль за ними. Однако очевидно, что внедрять потоковый ввод информации целесообразно только в случае больших объемов входящих документов.

Потоковый ввод документов подразумевает системный подход к организации процесса обработки документов, от их получения до сохранения извлеченных данных в хранилище (см схему 1). Результаты, достигнутые на промежуточных этапах, должны быть согласованы с последующими стадиями автоматизированного ввода. Потоковый ввод – это не просто набор программных и аппаратных инструментов, выстроенных в определенной последовательности, а проработанная система в организации всех составляющих элементов этого процесса, которая объединяет все операции в единую технологическую цепочку. Все документы, которые необходимо перевести в электронный вид, можно разделить на три категории.

Первая – это структурированные документы, к числу которых относятся заявления, жалобы, анкеты, т.е. любые унифицированные формы с фиксированной разметкой. Обычно такие документы применяются для сбора данных, например, это бланки, которые заполняются от руки. Отметим, что форму квитанции А7, предназначенную для ручного заполнения агентом, мож-

но привести к машиночитаемому виду, и тогда ее можно будет успешно распознавать.
Вторая категория документов – это гибкие формы для ввода данных, которые имеют значительно более сложный, слабоструктурированный вид. К их числу относятся счета-фактуры, накладные, акты выполненных работ, платежные поручения и прочее. Документы этого типа представляют, пожалуй, наибольшую сложность при вводе информации. Одновременно в них представлены цифры, буквы, графические элементы, причем все данные могут иметь различное расположение на странице. Качественный автоматизированный ввод информации в таких случаях способна обеспечить далеко не каждая программа, и адекватность полученных образов будет зависеть от обоснованного выбора программного инструмента. Характерным признаком третьей категории входных документов является наличие у них неструктурированного расположения информации. Это договоры, письма, служебные записки. Такие документы часто используются как доказательная или справочная информация. Поэтому ее хранение предусмотрено как в виде текста, удобного для выполнения поисковых задач, так и в графическом формате как подтвержденные копии юридически значимых документов.

Построение системы потокового ввода документов

Подготовка документов для потокового ввода осуществляется исходя из требований тех аппаратных средств, которые применяются на последующих этапах обработки. Перед сканированием документов надо удалить скрепки, проклеить отдельные страницы, устранить сгибы. Затем документы направляются на сканер, где производится их сканирование с последующим контролем качества полученных электронных копий. Операция может проводиться централизованно или удаленно на рабочих станциях в филиалах.

Следующий этап обработки связан с распознаванием полученных электронных копий. Для этих целей применяются специализированные программы, такие, как ABBYY FlexiCapture. Для выявления содержательной информации на отсканированных копиях применяются заранее описанные схемы (шаблоны). Программа определяет, какого вида получен документ, и находит на нем поля, соответствующие данному виду. Поля могут содержать разнообразную информацию: текст, метки, штрих-код или графическую картинку. Например, в заявлении на полис ОСАГО программа может распознать номер, дату, контактную информацию о страхователе, сумму страховки и другие данные.

Программа не только распознает содержимое полей, но и приводит их значения к заданному формату, если требуется. Применяются правила проверки, заложенные в шаблон (соответствие суммы в цифровой форме ее записи прописью, попадание даты в заданный интервал, проверка по внешней базе данных и т.п.).

Еще один этап потокового ввода – верификация полученных данных или, другими словами, подтверждение правильности проведенного распознавания. Часть проверки выполняется в автоматическом режиме для исправления типовых ошибок, например, замены десятичной точки на запятую, орфографическая коррекция по словарю и т.д. Специализированная часть процедуры верификации выполняется по специально разработанным сценариям. Страховщик заранее определяет, какие данные он хотел бы извлечь из документов, какие применить к ним правила проверки, какие поля требуют обязательной верификации, а какие нет. Заданий на проверку может быть много, и все они закладываются в шаблон документа. Настройку шаблона может выполнять как заранее обученные сотрудники страховой компании (для этого в ABBYY FlexiCapture есть все необходимые инструменты), так и компания-интегратор – все зависит от желания заказчика.

Отсканированный, распознанный и верифицированный документ передается в целевую систему (электронный архив, систему электронного документооборота, корпоративную информационную систему) как изображение с присущим ему уникальным набором данных (тип документа, его атрибуты, системная информация).

В дальнейшем по этой информации можно будет легко отыскать нужный документ в корпоративной информационной системе.

При распознавании каждого элемента программа определяет ве-

роятность, с которой она совершила эту операцию. Если вероятность распознавания оказывается ниже заданного порога, то символ помечается как «распознанный неуверенно», и право окончательного выбора передается человеку. Оператор должен подтвердить правильность введенного знака либо внести исправления. Если требуется очень высокая точность распознавания определенных данных в документе, то проверку можно делегировать оператору в обязательном порядке – вне зависимости от того, насколько в них «уверена» сама программа. В этом случае последнее слово всегда будет за человеком.

Кроме верификации распознанных символов, применяются логические правила проверки. Они позволяют отследить ошибки распознавания, а также опечатки и ошибки при заполнении документов. В случае остановки из-за непройденной проверки на ошибки документ возвращается на повторную обработку. Если же все проверки завершены успешно, то система переходит к перемещению данных в электронное хранилище. В результате создается единая база данных с электронным архивом, на которых строится фундамент для всей системы электронного документооборота в компании.

Потоковый ввод может быть реализован двумя способами: по индивидуальной или централизованной схеме. Индивидуальный вариант применяется для входящих документов, обрабатываемых «по мере поступления» непосредственно в точке приема документов. В компании создаются специальные рабочие места, где устанавливаются специализированные программы, такие как ABBYY FlexiCapture. Там же устанавливается сканер, любой модели, вплоть до самых дешевых планшетных аппаратов. Входящие документы обрабатываются сразу же после их приема у клиента. Время обработки одного листа формата А4 занимает обычно не более 1 минуты.

Второй вариант потоковой обработки относится к централизованному типу. При этом в точке приема документы только сканируются и пересылаются в центр для дальнейшей обработки. Далее в центре происходит распознавание документов и их верификация. Если нет возможности установить в точках приема сканеры, то бумажные документы отвозят в центральный офис, где их централизованно сканируют на высокопроизводительных промышленных сканерах. При таком подходе применяется принцип конвейера. За каждый этап (сканирования, распознавания, верификации, экспорта) об-

работки отвечает специальная станция или сервер, а на сканировании и верификации задействованы отдельные сотрудники. Если документооборот выполнять вручную, то при таких объемах неизбежно будут возникать значительные потери рабочего вре-

Пример проекта в «Цюрих. Ритейл»

В качестве примера реализации электронного документооборота можно рассказать о системе, созданной для страховой компании «Цюрих. Ритейл». Компания имеет развитую структуру региональных филиалов с большим объемом бумажных документов в работе. Ежедневно через нее проходит в среднем около 500 новых полисов ОСАГО, 160 договоров каско, 150 документов по другим видам страхования. Общий объем хранилища превышает 3 млн документов.

мени высококвалифицированных специалистов. Они будут вынуждены неэффективно тратить его на поиск, сбор и копирование документов. Например, отсутствие доступа к оригиналам документов из ее территориальных офисов часто становилось серьезным препятствием для быстрого обслуживания клиентов на местах. Решение было найдено в переходе на электронный документооборот с потоковым вводом документов. В Москве и региональных филиалах «Цюрих. Ритейл» были оборудованы более 100 рабочих мест, где установлено про-

граммное обеспечение Citrix, с помощью которого были созданы виртуализованные рабочие места с загрузкой лицензионного ПО с центрального сервера в Москве. Здесь же была установ-

лена и система для потокового ввода ABBYY FlexiCapture 8.0 Professonal, которую сотрудники территориальных офисов «Цюрих.Ритейл» стали использовать в режиме терминального доступа. Такая организация технологического процесса позволила обеспечить корректный ввод данных на местах, сохранив ответственность за качество ввода на сотрудниках региональных офисов. В центральном офисе в Москве решались задачи, связанные с администрированием системы в целом. Вся обобщенная информация и документы в электронном виде собирались здесь, а затем передавались в систему электронного документооборота IBM FileNet P8.

Как отметили в страховой компании, новая система смогла полностью обеспечить потребности компании в автоматизации ввода документов в электронный архив и позволила существенно снизить время на обслуживание клиентов. При этом было выбрано решение с использованием технологии виртуализации, надежность которого уже была подтверждена на практике многочисленными аналогичными внедрениями по всему миру.

Проект позволил компании «Цюрих.Ритейл» обеспечить размещение в системе электронного документооборота до 500 тысяч страниц клиентских документов ежемесячно, загрузив в работу все 100 территориально распределенных филиала. Полученный опыт показал: внедрение потоковой автоматизированной системы документооборота в страховой компании – это возможность предоставить сотрудникам надежный инструмент для работы, а для клиентов – получить качественный сервис при минимальных затратах времени на обслуживание.