

Рынок систем обработки деловых документов. Перспективы и направления развития

Д. Л. Шоломов, В. В. Постников, Н. Н. Никольский

В статье рассматриваются тенденции и перспективы развития систем массового ввода структурированных документов. Приведены оценки роста рынка систем данного класса, представлены основные фирмы-производители и указаны базовые направления развития систем. Одной из важных тенденций является наложение секторов бумажного и электронного оборота документов. В связи с этим в работе описана система массового ввода документов Cognitive Forms, позволяющая работать одновременно как с бумажным, так и с электронным представлением документа.

В настоящее время в сфере информационных технологий наблюдается динамичное развитие автоматизированных систем связанных с обработкой деловой информации. К данному классу относятся системы электронного документооборота, системы массового ввода структурированных документов, системы управления корпоративным контентом и другие.

По подсчетам экспертов Forrester Research, к 2008 году объем мирового рынка систем управления корпоративным контентом составит \$3,9 млрд, а объем рынка систем электронного документооборота достигнет уровня \$1,2 млрд, что составляет около трети объема рынка ЕСМ-систем, при этом на данный момент объем мирового рынка СЭД составляет около \$1,1млрд.

Что касается российского рынка, то его объем в настоящее время составляет \$1.2–\$1.6 млрд, при динамике роста порядка 25 % в год, по данным CNews Analytics, объем российского рынка только систем электронного документооборота в денежном выражении по итогам 2006 года превысил 140 млн долл. Ожидается, что в 2007 году он увеличится более, чем на \$30 млн и превысит \$170 млн. Согласно исследованию Strategy Partners Market Report динамика роста в этом секторе сохранится. Причинами подобного роста, по нашему мнению, является высокий потенциал рынка. По данным Gartner Group компании (потенциальные заказчики) тратят ежегодно на подготовку к печати и заполнение форм

порядка \$6 млрд и \$360 млрд на обработку форм. Таким образом, рынок еще далек от стадии насыщения.

Одной из основных причин, сдерживающих процесс внедрения электронного документооборота на российских предприятиях и в органах государственной власти, является сложность в использовании информации, содержащейся внутри неструктурированного документа, в управленческих целях. Система электронного документооборота технически не обладает возможностью собрать однородные данные, «разбросанные» в произвольном порядке внутри документов, а, следовательно, не способна эффективно осуществлять сбор и анализ информации, что значительно снижает ее функциональные возможности.

Организовать эффективную систему управления на предприятиях на основе современных информационных технологий было бы гораздо проще, если бы внутренние документы предприятия оформлялись бы по структурированной форме, регламентирующей расположение типовых данных в документе. Сделать это весьма несложно, поскольку, как показывает практика, более 70 % документов в общей структуре документооборота могут быть структурированы.

В первую очередь, это выгодно самим предприятиям, поскольку они получают возможность на основе внедренных информационных систем осуществлять экономическую оценку своей деятельности, что позволит снизить издержки и повысить норму прибыли за счет оптимизации внутренних бизнес-процессов.

Увеличение доли структурированных документов (форм) в документообороте неизбежно приведет к упрощению интеграции систем ввода и хранения форм со всевозможными программными приложениями (базами данных, электронными таблицами, интернет-технологиями, OLAP-средствами и т. д.). Таким образом, мы можем говорить о расширении понятия системы документооборота форм до системы структурирования управленческой информации. В результате использования подобных систем, заказчик практически мгновенно сможет получать статистические картины по собранным в формах данным, строить прогнозы, и максимально эффективно использовать полученные результаты для собственного бизнеса.

В нашей повседневной жизни мы постоянно сталкиваемся с документами, содержащими в себе набор полей или, проще говоря, формами. Это могут быть анкеты и квитанции, бланки заказа и накладные. Традиционные бумажные формы, заполняемые от руки или с помощью печатающих устройств, в настоящее время часто заменяются на электронные. На рис. 1 приведены примеры бумажных форм — платежное поручение ЦБ РФ, анкета застрахованного лица (форма АДВ-1) Пен-

The image shows three overlapping paper forms. The top-left form is a 'ПЛАТЕЖНОЕ ПОРУЧЕНИЕ № 5' (Payment Order No. 5) dated 19.01.98, for the sum of 200 rubles. The middle form is a 'АНКЕТА УЧАСТНИКА КОНФЕРЕНЦИИ' (Conference Participant Survey) from Cognitive Technologies Ltd. The bottom-right form is a 'АНКЕТА ЗАСТРАХОВАННОГО ЛИЦА' (Survey of Insured Person) for a person named Novikov Grigoriy Sergeevich, born 12 December 1980, in Voronezh, Russia.

Рис. 1. Примеры бумажных форм

сионного Фонда РФ, предназначенная для рукопечатного заполнения и анкета участника конференции.

К электронным формам относятся, например, форма заказа товара в интернет-магазине, или анкета, присланная для заполнения по электронной почте, на рис. 2 приведен пример одной и той же электронной формы в виде PDF и HTML. Однако бумажные формы не вытесняются электронными, скорее электронные формы позволяют расширить круг пользователей их заполняющих, а также ускорить процесс получения данных, особенно при использовании средств Интернет для пересылки форм. Бумажный документ по-прежнему используется в случае необходимости простановки печати и подписи.

Системы, призванные обеспечивать ввод и документооборот бумажных и электронных форм называют системами Forms processing. Наиболее удачным переводом Forms processing на русский язык является термин **Формооборот**.

Основными игроками на мировом рынке Forms processing являются компании:

- **Adobe + Accelio**¹⁾ — продукты Adobe Acrobat, Accelio Capture Form Flow, FormFlow Classic, Accelio Capture ReachForm;
- **Cardiff-TeleForm**, LiquidOffice, LiquidPDF;
- **Kofax (Dicom Group)** — Ascent Capture;
- **ScanSoft** — OmniForm;

¹⁾ В феврале 2002 г. было объявлено о приобретении Adobe Systems компании Accelio за \$72 млн.

Анкета участника конференции

Фамилия (рус.): Иванов

(англ.): Ivanov

Имя (рус.): Иван

(англ.): Ivan

Отчество: Иванович

Страна: РОССИЯ + Город: Москва

Форма участия: доклады на семинарах

Нужна ли Вам гостиница? ☒

Нужна ли Вам виза? ☐

Отправить Отмена

Рис. 2. Примеры электронной формы в виде PDF и HTML

- **ReadSoft** — DOCUMENTS;
- **AnyDoc** — продукты AnyDoc®CAPTUREit, AnyDoc®MANAGEit, AnyDoc®EXCHANGEit, AnyDoc®BROKERit.

Они занимают порядка 20–30 % рынка. Их обороты по данному направлению находятся в диапазоне от \$140 млн до \$500 млн. Другие компании работают в конкретных областях (государственный сектор, медицина, страхование, банковский и т. д.) и имеют обороты порядка \$7–\$10 млн. Все основные участники этого рынка увеличивают объем бизнеса не менее чем на 20 % (у Adobe) ежегодно. Например, динамика роста у компании Dicom Group в 2006 г. составила порядка 30 % по сравнению с предыдущим годом.

Следует отметить, что сегодня на мировом рынке систем класса Forms processing имеет место ярко выраженная тенденция по наложению секторов бумажного и электронного оборота форм, см. рис. 3. Становится все больше компаний, имеющих в своем арсенале программные средства из обоих сегментов. Далее, говоря о рынке Forms processing, мы будем иметь в виду именно сектор решений для документооборота одновременно электронных и бумажных форм.

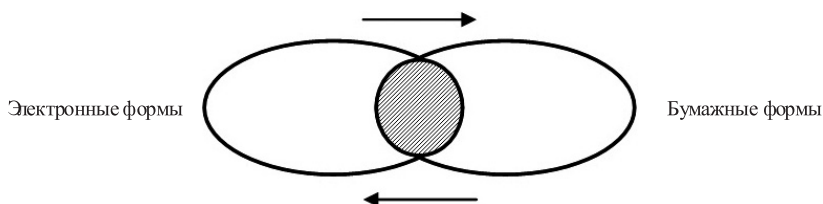


Рис. 3. Наложение секторов бумажного и электронного оборота форм

Российский рынок, по нашему мнению, подобный бум ожидает в ближайшие годы. Технологии обработки одновременно бумажных и электронных форм сильно востребованы среди потенциальных заказчиков. На это есть объективные причины. Количество бумажных форм в ближайшее время вряд ли уменьшится. Более того, по статистике их объемы растут из года в год. Это видно, в частности, на примере Пенсионного Фонда России. Что касается электронных форм, то потребность в их применении становится все более значимой в связи с массовым развитием средств электронных коммуникаций. Также следует отметить, что толчок к развитию систем электронного оборота и хранения документов был обусловлен тем, что в 2002 году начал действовать Федеральный закон «Об электронной цифровой подписи», а в 2006 году увидел свет Федеральный закон «Об информации, информационных технологиях и о защите информации». По нашему мнению, в ближайшие 5 лет российский рынок формооборота сможет достигнуть ста миллионов долларов.

Базовым решением для массового ввода бумажных документов от компании Cognitive Technologies является продукт **Cognitive Forms Network**, включающий набор основных и вспомогательных модулей. Основные модули соответствуют этапам обработки документа. Это станция сканирования **CF ScanPack**, сервер идентификации, сортировки и распознавания **CF Processor**, станция верификации **CF Editor**.

Основные модули развертываются в локальной сети и образуют технологическую линию, на входе которой — бумажные документы, на выходе — их электронные аналоги, готовые к загрузке в информационную систему предприятия, см. рис. 4. «Узкие места» конвейера расширяются увеличением числа рабочих станций на перегруженном этапе.

Вспомогательные модули реализуют дистанционное управление маршрутами и трассировку документов, сбор статистики (для расчета зарплаты операторов и для анализа работы комплекса и его компонент), протоколирование событий и операций, а также ряд других функций, обеспечивающих администрирование работы системы.

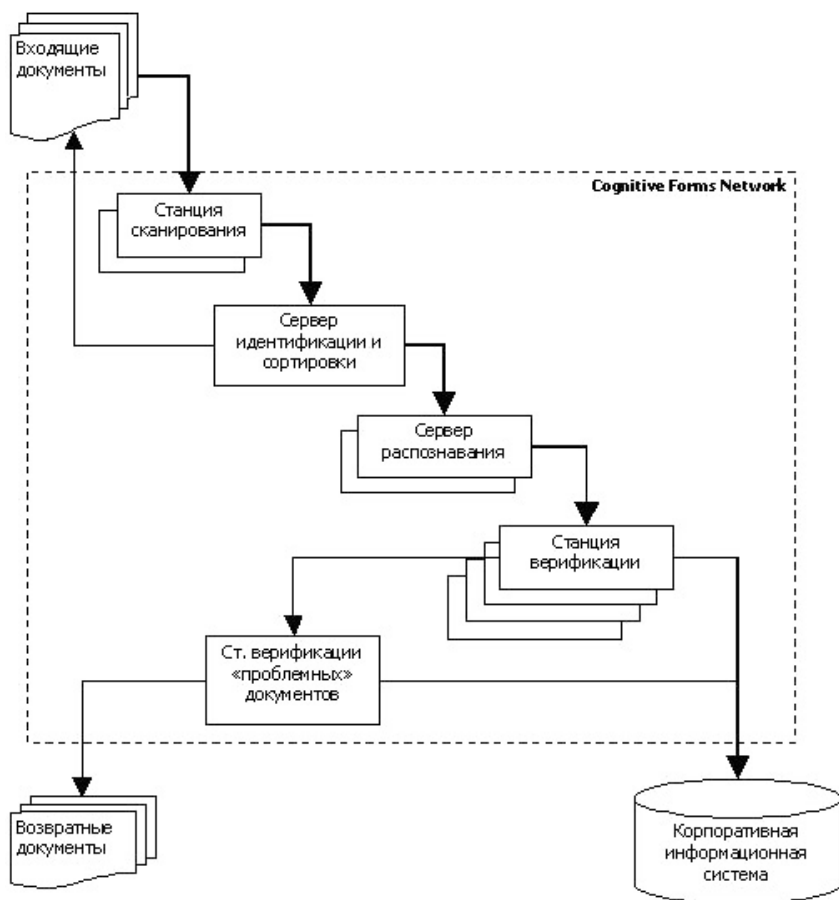


Рис. 4. Схема процесса ввода документа системой Cognitive Forms Network

Запуск типового проектного решения по вводу документов на основе **Cognitive Forms Network** включает этапы типографского макетирования машиночитаемых форм документов, настройки шаблонов распознавания, разработки и подключения системы логического контроля. Для проведения этих работ используется **Cognitive Forms Developer Kit**. Настроенная технологическая линия проходит обязательные стендовые испытания на тестовом наборе документов. Срок запуска типового проектного решения составляет 6–8 недель. В зависимости от сложности документов, объем тестового набора составляет несколько десятков или сотен документов, реалистичных по содержанию и качеству заполнения.

Распознающее ядро системы, входящее в базовую комплектацию продукта обучено на выборке более 7 млн реальных образцов печатных и рукописных документов и обеспечивает высокое качество распознавания и идентификации на широком спектре документов. Тем не менее, при организации ввода сложных документов в виде «как есть» (не оптимизированном для машинной обработки) для улучшения качества распознавания иногда требуется дополнительное обучение и специализированная адаптация распознающего ядра системы. Объем тестового набора в этом случае может составлять несколько тысяч или десятков тысяч документов.

В 2003–2006 годах компанией Cognitive Technologies был реализован ряд крупных проектов по массовому вводу документов. В основе проектных решений лежит программный комплекс **Cognitive Forms Network**, настроенный на распознавание и ввод деловых документов. Например, на ОАО «Магнитогорский Металлургический Комбинат» в рамках систем запущенных в промышленную эксплуатацию, вводятся счета-фактуры, договоры, спецификации, акты и накладные, товаросопроводительные документы — исходящие счета фактуры, сертификаты качества, ж/д квитанции, дубликаты накладных и другие документы. Сканирование осуществляется как при помощи промышленного сканера Kodak i620, так и при помощи персональных документных сканеров типа Fujitsu fi-4220C2 в нескольких подразделениях предприятия. Программный комплекс интегрирован с Корпоративной Информационной Системой, реализованной на базе Oracle E-Business Suite и электронным архивом на базе IBM Content Manager. Схема интерфейса взаимодействия системы по вводу договоров и отгрузочных разрядок АСВ ДиОР приведена на рис. 5.

Нужно подчеркнуть, что вводимые документы, такие как договоры, спецификации, являются слабоструктурированными. При определении типа документа и локализации необходимых для распознавания реквизитов, используются алгоритмы так называемого атрибутивного разбора. При этом на специальном языке структура документа описывается не полностью, а описанию подлежат только фрагменты, содержащие информацию непосредственно об интересующих атрибутах. Кроме того, при вводе ряда документов используются современные методы обработки изображений — адаптивная фильтрация шумов и бинаризация. Необходимость их использования вызвана невысоким качеством документов, например, регистрационные документы контрагентов поступают в виде ксерокопий с «гербовой» бумаги. Отдельное внимание уделено описанию полей ввода при помощи порождающих грамматик и синтаксическим методам их контекстной обработки.

Программные комплексы, внедренные на ряде российских предприятий, лежат в русле мировой практики. Например, в настоящее время

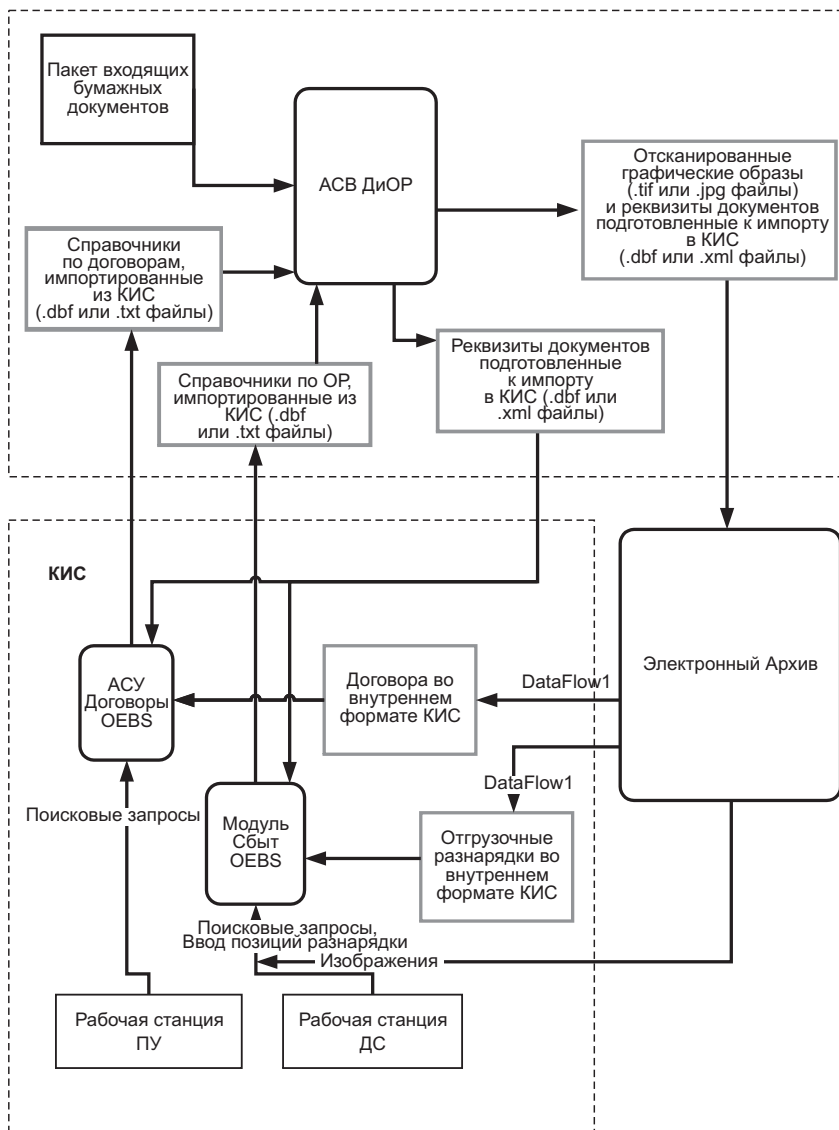


Рис. 5. Схема интерфейса взаимодействия АСВ ДиОР

существуют десятки систем по вводу счетов-фактур, которые широко используются в рамках автоматизации предприятия. Наиболее известными являются:

- **ReadSoft:** DOCUMENTS for INVOICES;
- **Kofax:** Ascent for Invoices;
- **Cardiff:** Cardiff Teleform v. 10;
- **AnyDoc:** AnyDoc®INVOICE.

По данным компании ReadSoft автоматизация процесса ввода счетов-фактур позволяет экономить не менее \$100 000 в год при объеме 20 000 документов. Учитывая то, что ориентировочный объем поступающих счетов-фактур на крупное предприятие составляет 300 000 в год, размер прибыли составляет \$1 500 000. Оценка затрагивает только непосредственно процесс ввода и не включает значительно большую экономию средств, получаемую от автоматизации корпоративного документооборота.

Следует отметить, что перечисленные системы по вводу счетов-фактур не позволяют вводить российские счета-фактуры, которые являются существенно менее структурированными в виду более слабой их стандартизации. Фактически на данный момент компания Cognitive Technologies является лидером на российском рынке по вводу слабоструктурированных многостраничных документов.

Как уже было сказано, в мире наметилась устойчивая тенденция к объединению бумажных и электронных документов. Мировые лидеры выпускают решения для работы, как с бумажными, так и с электронными документами. Например, у компании Cardiff система Cardiff Teleform предназначена для ввода бумажных документов, а Cardiff LiquidOffice eForms — для электронных. Компания Cognitive Technologies, идя в ногу с мировыми производителями, представляет на российском рынке комплексное решение, позволяющее осуществлять одновременную обработку как бумажных, так и электронных документов. Данная технология расширяет возможности продукта **Cognitive Forms Network**.

Условно технологическая схема состоит из трех уровней, см. рис. 6—8. Первый уровень отвечает за дизайн макета формы. Центральным модулем первого уровня является **CF Designer**, позволяющий создавать макет документа на основании бумажного либо электронного образца. При дизайнировании используются алгоритмы векторизации исходного изображения. Созданный макет бланка можно распечатать на принтере, при этом возможно проводить частичную надпечатку ряда полей на основе предопределенного списка значений из БД. Также макет можно сохранить в виде электронной формы в формате HTML и PDF.

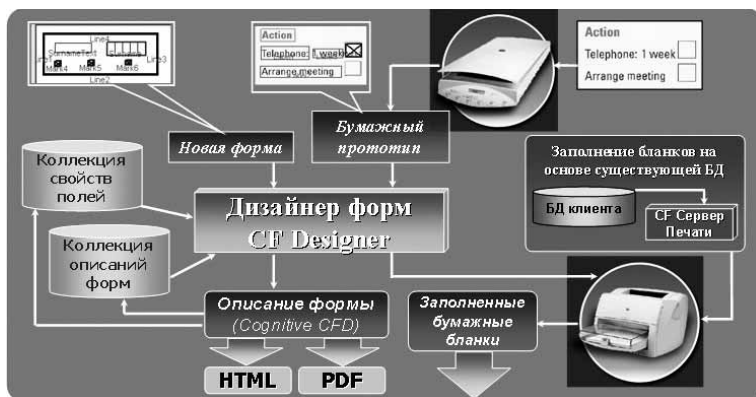


Рис. 6. Уровень дизайна и публикации документа

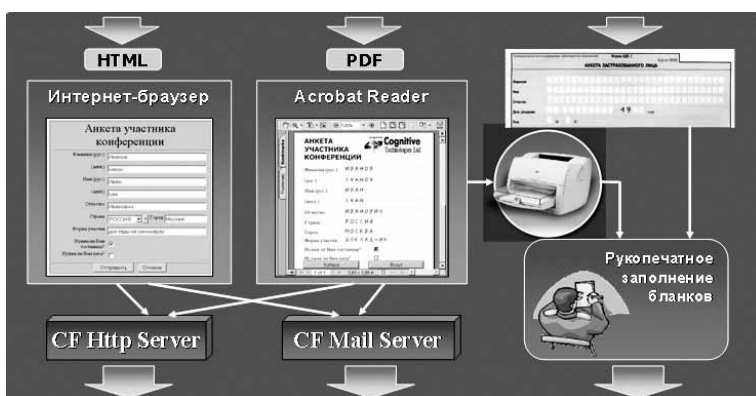


Рис. 7. Уровень средств заполнения документа

Второй уровень отвечает за внесение данных в поля документа и их пересылку на сервер для дальнейшей обработки. При этом бумажные бланки либо заполняются вручную, либо значения впечатываются при помощи принтера. HTML документы заполняются из браузера, далее введенная информация пересылается на HTTP или почтовый сервер. PDF документ заполняется либо в электронном, либо в бумажном виде.

Третий уровень отвечает за сбор данных, при этом данные, поступившие в электронном виде на **CF Http Server** и **CF Mail Server**, агрегируются в едином хранилище при помощи модуля **CF Data Collector**. А информация с бумажного носителя проходит стандартный цикл обработки (сканирование, распознавание и верификацию), а затем также как

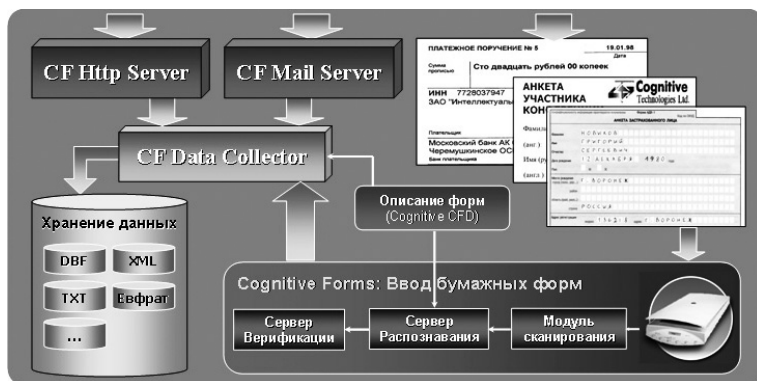


Рис. 8. Уровень сбора и хранения данных

и в случае с электронными документами, передается модулю **CF Data Collector**. При помощи модулей экспорта собранная в едином хранилище информация целиком либо порциями экспортируется во внешние информационные системы посредством специальных форматов и протоколов обмена данными.

Использование описанной технологии дает следующие существенные преимущества. К технологическим преимуществам можно отнести возможность быстрого моделирования форм средствами модуля **CF Designer**, наличие средств интеграции в виде отдельных компонентов экспорта и контроля целостности документа. Также следует отметить, что система позволяет проводить удаленный документооборот посредством использования электронных форм. К управленческим преимуществам можно отнести возможность качественной работы с информацией, наличие средств быстрой настройки системы под меняющиеся управленческие задачи и наличие средств удаленного управления.

Литература

1. Арлазаров В. В., Постников В. В., Шоломов Д. Л. Cognitive Forms — система массового ввода структурированных документов // В сборнике «Управление информационными потоками». М.: УРСС, 2002. С. 35–46.
2. Арлазаров В. В. Управление информационными потоками в системе автоматического ввода документов // Управление информационными потоками / Сборник трудов ИСА РАН. М.: УРСС, 2002.
3. Арлазаров В. Л., Емельянов Н. Е. Системы обработки документов. Основные компоненты // Управление информационными потоками / Сборник трудов ИСА РАН. М., УРСС. 2002.

4. *Postnikov V. V., Sholomov D. L., Marchenko A. E.* FlexiDocs: The Template Driven Document Recognition Technology // Proceedings of the 6th German—Russian Workshop on Pattern Recognition and Image Understanding (OGRW-6). 2003.
5. *Постников В. В., Марченко А. Е., Шоломов Д. Л.* Разбор структурированного документа в модели с нечеткой логикой // Сборник трудов ИСА РАН «Документооборот. Концепции и инструментарий». М.: УРСС, 2004. С. 71–82.
6. *Постников В. В.* Формальный подход к задаче идентификации графических образов структурированных документов // В сборнике «Развитие безбумажных технологий в организационных системах». М., 1999.
7. *Постников В. В.* Автоматическая идентификация и распознавание структурированных документов // Дис. ... канд. технич. наук. М., 2001.
8. *Sholomov D. L.* Syntactical Approach to Post-Processing of Fuzzy recognized Text // Proc. of The International Conference on Machine Learning, Technologies and Applications. CSREA Press, June 2003, USA. P. 115–121.