На правах рукописи

Завозкин Сергей Юрьевич

ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ИНТЕГРАЦИИ ИНФОРМАЦИОННЫХ СИСТЕМ НА ОСНОВЕ СИСТЕМЫ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА

05.25.05 — информационные системы и процессы, правовые аспекты информатики

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Кемерово — 2007

Работа выполнена в Кемеровском государственном университете на кафедре ЮНЕСКО по НИТ и Институте вычислительных технологий СО РАН

Научный руководитель: кандидат физико–математических наук,
доцент Гудов Александр Михайлович

Официальные оппоненты: доктор технических наук,
профессор Потапов Вадим Петрович

кандидат технических наук,
доцент Пестунова Тамара Михайловна

Ведущая организация: Томский государственный университет

Защита состоится 9 ноября 2007 г. в 10 часов 15 минут на заседании диссертационного совета Д 003.046.01 при Институте вычислительных технологий СО РАН по адресу: 630090, г. Новосибирск, просп. Ак. Лаврентьева, 6.

С диссертацией можно ознакомиться в специализированном читальном зале вычислительной математики и информатики ГПНТБ СО РАН.

Автореферат разослан 8 октября 2007 г.

Учёный секретарь
диссертационного совета
доктор физико-математических наук,
профессор

Л.Б. Чубаров

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Использование на предприятии системы электронного документооборота (СЭД) решает такие важные задачи, как автоматизация работы с документами и бизнес-процессами, обеспечение совместной работы с данными, обеспечение безопасности и надежности хранения информации. На данный момент на рынке существует большое число таких систем, обладающих, как правило, схожими базовыми функциями. Существенные различия между ними появляются лишь в реализации конкретных задач.

Главной проблемой таких систем является недостаточная развитость имеющихся средств интеграции с другими системами. Данная проблема весьма актуальная, так как на многих предприятиях существует большое число разрозненных информационных систем (ИС), постепенно приобретаемых или разрабатываемых в процессе жизни предприятия. Эти ИС зачастую обладают различными функционалом, логикой, архитектурой и форматом хранения данных. Причём большая часть таких систем создавалась разными разработчиками для решения определённых задач, а, следовательно, системы содержат лишь самые простые (на уровне передачи информации файлами определенной структуры) механизмы интеграции с другими ИС. Эта ситуация приводит к целому ряду проблем, таких как многократное дублирование хранимой информации, сложность поиска необходимых для синхронизации данных, низкая надёжность хранения и невысокая эффективность работы с данными, а также сложность поддержки целостности и непротиворечивости хранимых данных. Для решения перечисленных проблем необходима система, позволяющая выступить в качестве связующего звена при объединении ИС в одно информационное пространство. Наибольшая эффективность такого объединения достигается в том случае, если системы будут поддерживать несколько способов интеграции. Самый простой из них заключается в возможности использования одной и той же информации. Другой способ основан на использовании стандартизованного описания передаваемых данных и предоставлении системам набора сервисов для работы с ним Третий способ осуществляет связь ИС за счёт создания в связующей системе специальных бизнес-процессов (БП), позволяющих объединять внутренние БП этих систем. Как правило, СЭД содержат в себе механизмы интеграции, направленные на решение конкретных задач и поддерживают лишь некоторые из перечисленных способов интеграции.

При интеграции ИС важную роль играет импорт и экспорт данных. Так как большинство СЭД основаны на описании документов с использованием метаданных, то при импорте документов актуальной становится задача автоматического определения метаданных. Отсутствие у СЭД такого механизма является ещё одной существенной проблемой.

Таким образом, актуальной является задача разработки системы, позволяющей на своей основе различными способами интегрировать существующие на предприятии ИС, а также обеспечивающей автоматизацию управления документами и бизнес-процессами. Очевидно, что такая система должна относиться к классу СЭД.

Цель работы - разработка комплекса моделей, обеспечивающих информационную поддержку интеграции информационных систем на основе СЭД, как промежуточного ПО, также автоматизирующей управление документами и бизнес-процессами.

Задачи исследования

1. Провести анализ существующих подходов к интеграции ИС. Определить основные принципы и требования к модели комплексной интеграции ИС. Провести анализ популярных СЭД на предмет использования их в целях интеграции.
2. Разработать модель интеграции ИС на основе СЭД, обеспечивающей поддержку информационно-ориентированного, сервисно-ориентированного и процессно-ориентированного принципов интеграции. Разработать требования к моделям СЭД.
3. Разработать комплекс моделей, обеспечивающих информационное обеспечение интеграции. Провести верификацию на соответствие требованиям к моделям.
4. Разработать метод выбора оптимальной архитектуры СЭД на основе решения задачи оптимизации стоимости документопотоков.
5. Разработать модель автоматического определения метаданных электронных документов.
6. Реализовать СЭД и провести её апробацию в процессе управления ВУЗом.

К **объектам исследования** относятся методы интеграции информационных систем, электронные документы, документопотоки, процессы движения и управления электронными документами, бизнес-процессы и механизмы управления ими.

Методы исследования определяются сущностью теоретических и практических проблем. Используются: структурный системный анализ, теория баз данных, теория автоматизации проектирования, методы функционального проектирования в нотациях IDEF0, IDEF3 и IDEF1X, методы объектно-ориентированного проектирования, методы оптимизации, теория нейронных сетей, теория и методы разработки программного обеспечения, методы тестирования программного обеспечения.

Научная новизна

1. Предложена модель комплексной интеграции, позволяющая создавать системы, объединяющие в единое информационное пространство разрозненные ИС.
2. Предложены информационная, функциональные модели и модель структуры данных, позволяющие создавать на их основе системы, реализующие модель комплексной интеграции ИС, а также автоматизирующие управление документами и бизнес процессами.
3. Предложен метод выбора архитектуры СЭД на основе решения задачи оптимизации стоимости документопотоков.
4. Предложен метод автоматического определения метаданных электронного документа, в основу которого положена математическая модель нейронной сети Кохонена.
5. Реализована СЭД, обеспечивающая автоматизацию управления документами и бизнес-процессами, а также содержащая механизмы интеграции информационных систем.

Положения, выносимые на защиту

1. Модель комплексной интеграции ИС, обеспечивающая поддержку информационно-ориентированного, сервисно-ориентрованного и процессно-ориентированного принципов интеграции.
2. Информационные объекты системы, информационная и функциональные модели, а также модели структуры данных СЭД.

3. Метод выбора архитектуры СЭД на основе решения задачи оптимизации стоимости документопотоков
4. Метод автоматического определения метаданных документа, в основу которого положена математическая модель нейронной сети Кохонена.
5. Реализация СЭД, базирующейся на построенных моделях и позволяющей интегрировать на своей основе информационные системы, а также автоматизировать управление документами и бизнес-процессами.

Практическая ценность результатов работы заключается в разработанных моделях и методах, составляющих основу СЭД. Они могут быть использованы при построении сложных систем, обладающих схожим функционалом, и, в первую очередь, систем, обеспечивающих автоматизацию деятельности предприятия и интеграцию ИС.

Основные результаты работы были использованы при выполнении следующих проектов:

- Грант Министерства Образования и Науки Российской Федерации № 4828 в рамках федеральной программы «Развитие научного потенциала высшей школы» (2005 – 2006 год).
- Госконтракт № 12/10 АКО на выполнение работ по мероприятию “Создание областного реестра информационных ресурсов, баз данных научно-технической информации, информации учебного назначения и электронных средств обучения в учреждениях НПО, СПО, ВПО, ДПО” (2006 год).
- Грант Министерства Образования и Науки Российской Федерации № 4256 “Создание типового информационно-вычислительного портала для организации учебной и научной деятельности ВУЗа” в рамках аналитической ведомственной целевой программы “Развитие научного потенциала высшей школы (2006-2008 годы)” (2006-2007 годы).
- Грант Ученого Совета КемГУ “Система электронного документооборота ВУЗа”. Протокол заседания учёного совета КемГУ № 3 от 24.03.2007 (2007 год).

Обоснованность и достоверность полученных результатов подтверждается реализацией системы электронного документооборота, её внедрением в опытную эксплуатацию в Кемеровском государственном университете (http://sed.kemsu.ru).

Апробация результатов. Основные результаты диссертации представлялись на следующих конференциях: Международной научно-практической конференции “Новые информационные технологии в университетском образовании” (Кемерово 2006), Международной конференции “Вычислительные и информационные технологии в науке, технике и образовании” (Павлодар, 2006), Международной конференции по электронным публикациям “El-Pub 2003” (Новосибирск, 2003), Международной научно-практической конференции “Информационные технологи и математическое моделирование” (Анжеро-Судженск, 2006), Всероссийских научно-практических конференциях “Инновационные недра Кузбасса. IT-технологии”, “Недра Кузбасса. Инновации”, “Информационные недра Кузбасса” (Кемерово, 2007, 2006, 2005), Всероссийских научно-практических конференциях “Системы автоматизации в образовании, науке и производстве” (Новокузнецк, 2005, 2007), Всероссийской конференции с участием иностранных учёных “Распределённые информационно-вычислительные ресурсы” (Новосибирск, 2005), Всероссийской конференции молодых ученых по математическому моделированию и информационным технологиям (с участием иностранных ученых) (Кемерово, 2005).

Публикации

По теме диссертации опубликовано 23 работы, в том числе (в скобках в числителе указан общий объём этого типа публикаций, в знаменателе – объём, принадлежащий лично автору) 4 статьи в изданиях, рекомендуемых ВАК для предоставления основных результатов диссертации (2,06/1,78 печ. л.), 1 статья в научном журнале (0,5/0,43), 11 публикаций в трудах и материалах конференций (3,5/2,73 печ. л.), 7 публикаций в тезисах конференций (0,93/0,81 печ. л.).

Личный вклад автора. Основные научные и практические результаты диссертации получены автором лично. Из печатных работ, опубликованных диссертантом в соавторстве, в диссертацию вошли только те результаты, которые автором получены лично на всех этапах: от постановки задач и моделирования, до реализации системы.

Структура и объём работы. Диссертационная работа состоит из введения, трёх глав, заключения и списка литературы. Общий объём работы составляет 168 страниц машинописного текста, включая 40 иллюстраций, 6 таблиц и библиографический список из 116 литературных источников. В работе имеется 5 приложений объёмом 50 страниц.

Автор выражает глубокую благодарность и признательность д.ф.-м.н., профессору К.Е. Афанасьеву за постоянное внимание, многочисленные обсуждения и ценные замечания, способствовавшие успешному выполнению данной работы, а также за помощь и поддержку в процессе выполнения диссертационной работы.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассматриваются наиболее распространенные на отечественном рынке программного обеспечения СЭД, их достоинства и недостатки. Обосновывается актуальность исследования построения СЭД, обеспечивающей автоматизацию управления документами и интеграцию на её основе различных ИС. Определяется направление исследований настоящей работы и цели, поставленные перед автором диссертации.

Первая глава посвящена анализу предметной области исследуемой проблемы. Результаты анализа подходов к интеграции ИС и существующих СЭД показали актуальность разработки собственной СЭД и позволили сформулировать требования к моделям системы.

Можно выделить несколько уровней интеграции:

- *Интеграция бизнес-процессов* – основана на определении, реализации и управлении процессами обмена информацией между различными бизнес-системами.
- *Интеграция приложений* – основана на объединении данных или функций одного приложения с другим, благодаря чему обеспечивается интеграция, близкая к реальному времени.
- *Интеграция данных* – основана на идентификации и каталогизации данных с целью их дальнейшего использования.
- *Интеграция на основе стандартов* – основана на использовании стандартных форматов данных (например, CORBA, JavaRMI, XML).
- *Интеграция платформ* – касается процессов и инструментов, с помощью которых системы могут осуществлять безопасный и оптимальный обмен информацией.

Помимо этого принято выделять три принципа интеграции:

- *Информационно-ориентированный* - основан на использовании одной и той же информации двумя и более системами. При этом у каждой системы имеется набор от-

крытых сервисов для работы с информацией. Данный принцип реализуется на уровнях интеграции приложений.

- *Сервисно-ориентированный* - основан на использовании стандартизованного описания формата передачи данных. При этом данные хранятся в единой базе данных системы middleware и имеется набор сервисов для работы с ними. Данный принцип реализуется на уровнях интеграции данных, платформ и на уровне использования стандартов интеграции.
- *Процессно-ориентированный* – основан на возможности присоединения к внутренним прикладным процессам каждой ИС таким образом, чтобы не просто использовать их функции, а создать новый бизнес-процесс для связи этих ИС. Данный принцип реализуется на уровнях интеграции бизнес-процессов, платформ и на уровне использования стандартов интеграции.

При этом ни один из уровней или принципов интеграции не является универсальным, и не существует общего способа решения задачи интеграции ИС. В зависимости от ситуации наиболее удачным решением оказываются различные их комбинации. Поэтому максимальная эффективность достигается в том случае, если интеграция осуществляется на основе системы middleware, обеспечивающей все три принципа интеграции. В качестве такой платформы предлагается использовать СЭД, обладающую рядом дополнительных свойств и осуществляющую информационное обеспечение трёх вышеизложенных принципов. При этом далеко не каждая СЭД может быть для этого использована. Основными причинами этого являются:

- **Невозможность обеспечения в полной мере сервисно-ориентированного принципа интеграции.** Для обеспечения данного принципа СЭД должна предоставлять возможность хранения и обработки как структурированных, так и слабоструктурированных данных, снабжая их стандартизованным структурированным описанием, и устанавливая отношения между данными. Такая СЭД должна иметь набор интерфейсов, предоставляющих другим ИС возможность работы с этими данными.
- **Невозможность обеспечения в полной мере процессно-ориентированного принципа интеграции.** Для обеспечения данного принципа система должна обеспечивать возможность управления бизнес-процессами и предоставлять интерфейс другим ИС для создания в ней типовых бизнес-процессов и управления ими в своих целях.
- **Отсутствие возможности автоматического определения метаданных из содержимого текстового документа**, что необходимо для осуществления эффективного импорта слабоструктурированных данных в СЭД из других ИС. Этот механизм также полезен при потоковом вводе документов.
- **Отсутствие возможности выбора оптимальной архитектуры СЭД**, что в ряде случаев может привести к ограничениям функциональности системы в целом.

Результаты анализа 9 наиболее популярных на рынке программного обеспечения отечественных систем электронного документооборота (DocsVision, Albetty, DIRECTUM, ЕВФРАТ-Документооборот, OfficeMedia, ДЕЛО, PayDox, ЛЕТОГРАФ, Documentum) показали, что все рассмотренные системы делятся на две группы по тем ключевым объектам, с которыми они работают, и, как следствие, по назначению систем. К первой группе относятся системы, ключевыми объектами которых являются только документы. Ко второй группе - системы, ключевыми объектами которых являются документы и бизнес-процессы. Большая часть из рассмотренных систем относится ко второй группе.

Все рассмотренные СЭД обладают схожими базовыми функциями, а существенные различия появляются в реализации конкретных механизмов. Это, например, средства расширения функционала систем, графические редакторы для создания БП, редакторы отчётов, развитые средства поиска и т.д. Результатом анализа функционала стал полный список функций систем с указанием особенностей их реализации.

Архитектура систем, по большей части, является трехзвенной. Обычно в СЭД для хранения данных используются базы данных как собственной разработки (например, БД компаний Гарант Интернейшенл и Cognitive Technologies), так и известных разработчиков (например, Oracle, MS SQL Server, Lotus Notes/ Domino, MS Access). Логика систем чаще всего располагается на сервере приложений, зачастую объединённом с web-сервером, а в некоторых случаях и с СУБД. Клиентская часть, как правило, устанавливается на рабочую станцию пользователя. Если предоставляется возможность работы из стандартного web-броузера, то пользователю доступен ограниченный функционал. Существует также ряд систем, ориентированных исключительно на web-клиента. При внедрении таких систем пользователь выбирает архитектуру из заранее определенного списка доступных решений.

К основным недостаткам рассмотренных систем можно отнести отсутствие возможности автоматического определения метаданных из содержимого текстового документа и недостаточная развитость имеющихся средств интеграции.

Анализ СЭД показал, что ни одна из рассмотренных систем не может быть использована в качестве платформы middleware, так, чтобы в полной мере обеспечить все три принципа интеграции.

В результате были предложены следующие требования к моделям СЭД. Модели должны обеспечивать:

1. Возможность эффективного добавления, редактирования, удаления и работы с данными произвольного (как структурированного, так и слабо структурированного) типа и размера.
2. Возможность совместного использования различными ИС одной и той же информации, которая должна осуществляться за счёт предоставления системам сервисов для работы с данными СЭД, и использования сервисов ИС для работы с их данными.
3. Возможность интеграции ИС за счёт использования стандартизованного описания формата передачи данных и наличия у систем схожих сервисов работы с ним.
4. Возможность объединения внутренних прикладных процессов каждой ИС за счёт создания в СЭД типовых бизнес-процессов.
5. Возможность автоматического определения метаданных из содержимого текстового документа, а также определения метаданных из описания, представленного в специальном формате (например, в XML формате).
6. Возможность управления документами (создание, редактирование, удаление документов, ведение истории работы с документами, установление ссылок между документами, автоматизация заполнения части метаданных, обеспечение совместного доступа к данным, управление версиями документов, вложение ссылок на документы, расположенные во внешних ИС).
7. Возможность маршрутизации (поддержка жесткой и динамической маршрутизации, мониторинг текущей деятельности пользователей с документами и заданиями, контроль текущего состояние документов).

8. Возможность управления БП, включая моделирование, выполнение, мониторинг и наличие интерфейса для создания БП.
9. Возможность реализации удобных средств навигации и организации доступа пользователей к информации (наличие отношений порядка между объектами в системе за счёт поддержки древовидной структуры представления данных, обеспечение эффективного поиска и управления данными, наличие закладки, позволяющей сохранять важные документы и контакты, для последующего быстрого доступа к ним).
10. Безопасность хранения и передачи информации (поддержка аутентификации, авторизации, разграничения прав доступа к объектам СЭД, делегирования прав доступа к документу).
11. Механизм настройки СЭД на выбранную архитектуру.

Во **второй главе** описываются: модель комплексной интеграции информационных систем, информационная и функциональные модели, модель структуры данных СЭД, метод автоматического определения метаданных из содержимого текстового документа, метод выбора архитектуры системы на основе решения задачи оптимизации стоимости документопотоков.

Для удовлетворения разработанным требованиям в информационную модель были введены следующие понятия объектного подхода:

- *Объект* – сущность с определёнными свойствами.
- *Класс объектов* – описание множества однородных объектов, имеющих одинаковые атрибуты, отношения с другими объектами и семантику.
- *Метаданное (атрибут)* – элемент, описывающий свойство объекта.
- *Отношение* – семантическая связь между объектами. Для описания отношений использована модель RDF.

Для удобства работы со схожими объектами введено понятие *группы объектов* – множество объектов, связанных отношениями определённого типа.

Результаты анализа существующих подходов к интеграции позволили ввести понятие *интеграции ИС* – объединение ИС, связывающее множество документов и отношений в данных системах. Под *ИС* понимается множество связанных различными отношениями документов, описывающих некоторые сущности (объекты, факты или понятия).

Определение ключевого объекта СЭД, электронного документа, разработано на основе ГОСТ Р 51141-98 “Делопроизводство и архивное дело. Термины и определения”, закона РФ N 24-ФЗ от 20.02.1995 “Об информации, информатизации и защите информации”, закона РФ от 10.01.2002 N 1-ФЗ "Об электронно-цифровой подписи" и ГОСТ Р ИСО 9000-2001: «Системы менеджмента качества. Основные положения и словарь». *Электронным документом* называется информационная пара: $d_i = \langle C_{d_i}, M_{d_i} \rangle$, обрабатываемая СЭД и циркулирующая в ней, где C_{d_i} – содержимое ЭД, M_{d_i} – метаданные ЭД.

На основе ГОСТ Р 51141-98. “Делопроизводство и архивное дело” разработана классификация электронных документов, включающая: наименование, тип, направление, способ наполнения, степень сложности, содержание, вид, срочность, степень гласности, срок хранения, стадия создания, род деятельности.

Под *содержимым* понимается информационное наполнение ЭД, которое представляется в виде вложенного файла или набора файлов произвольного типа и размера, а также в виде ЭД или набора электронных документов.

Метаданные ЭД – описание ЭД, однозначно его идентифицирующее, где отражаются как его статические, так и динамические характеристики. Структура метаданных ЭД разработана на основе *спецификации IMS* и стандарта метаданных *Дублинского ядра* с учётом разработанной классификации ЭД. При этом для каждого класса ЭД часть метаданных является обязательной. На основании разработанной структуры метаданных предложен XML формат их представления.

Ещё одним ключевым объектом СЭД является понятие бизнес-процесса. Для его определения введены понятия задания, функции и процесса. *Заданием* t_i называется множество: $t_i = \langle e_i, o_i, i_i, M_{t_i} \rangle$, где e_i – действие; o_i – объект СЭД, над которым выполняется действие; i_i – исполнитель, в роли которого может выступать ИС или пользователь СЭД, ответственный за выполнение действия; M_{t_i} – метаданные задания.

Функцией f_i называется пара $f_i = \langle e_i, M_{f_i} \rangle$, где e_i – элементарное действие, автоматически выполняемое системой и возвращающее определённый результат; M_{f_i} – метаданные функции.

Процессом p_i называется множество $p_i = \langle P, F, T_{d_i}, M_{p_i} \rangle$, с заданным регламентом выполнения элементов этого множества, где P – другие процессы; F – функции; T_d - связанные с ЭД d_i задания; M_{p_i} – метаданные задания.

Процесс связан с конкретным ЭД и описывает его маршрутизацию. *Маршрутом ЭД* называется последовательность связанных с ЭД заданий в соответствии с регламентом их выполнения. Маршруты ЭД подразделяются на жесткие (регламент выполнения заданий неизменен) и динамическими (регламент может быть изменён в процессе реализации маршрута, а также могут быть удалены или добавлены задания).

Бизнес-процессом b_i называется множество $b_i = \langle P, F, T, M_B \rangle$, с заданным регламентом выполнения элементов этого множества, где P – процессы; F – функции; T - задания; M_B – метаданные бизнес-процесса. Бизнес-процесс направлен на достижение определенной бизнес-цели.

Для обеспечения возможности управления безопасностью введены понятия *пользователя, права* и *роли.*

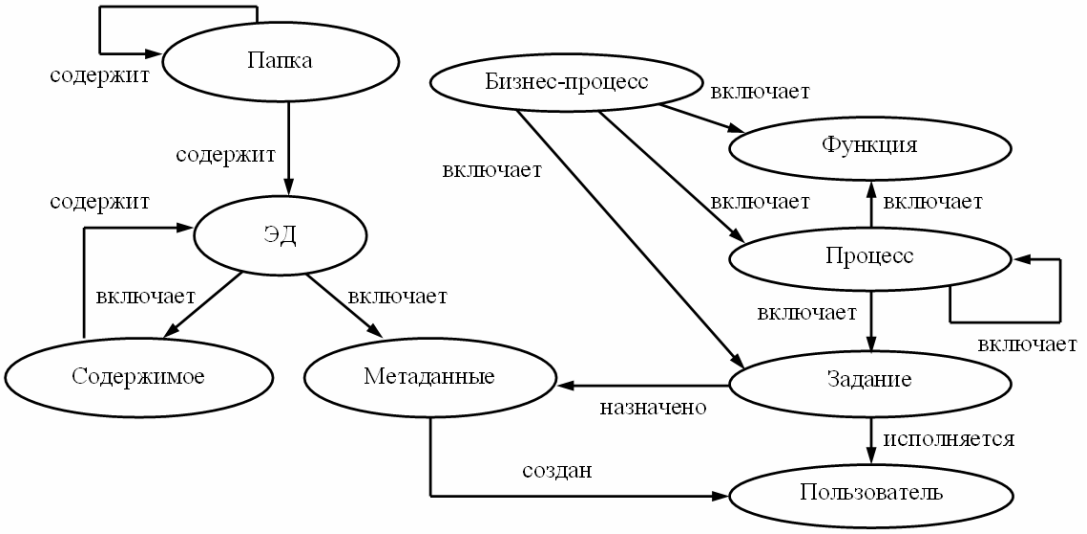

Рис. 1. Отношения основных объектов информационной модели

Для всех объектов информационной модели представлено табличное описание структуры метаданных и отношений с другими объектами. Отношения основных информационных объектов представлены на рис. 1.

Для описания функций системы, а также существующих в ней потоков данных построен комплекс функциональных моделей СЭД в стандартах IDEF0 и IDEF3.

Согласно разработанным требованиям к моделям СЭД и в соответствии с информационной моделью системы построена комплексная модель интеграции ИС. В модели применяется подход, заключающийся в использовании СЭД в роли системы middleware для интеграции ИС. СЭД обеспечивает поддержку трех основных принципов интеграции. Модель комплексной интеграции базируется на основных объектах и отношениях построенной информационной модели, включает набор моделей, каждая из которых соответствует одному из принципов интеграции.

В *модели информационно-ориентированной интеграции* используется единая база данных (БД) для хранения информации всех действующих и разрабатываемых ИС (см. Рис. 3).

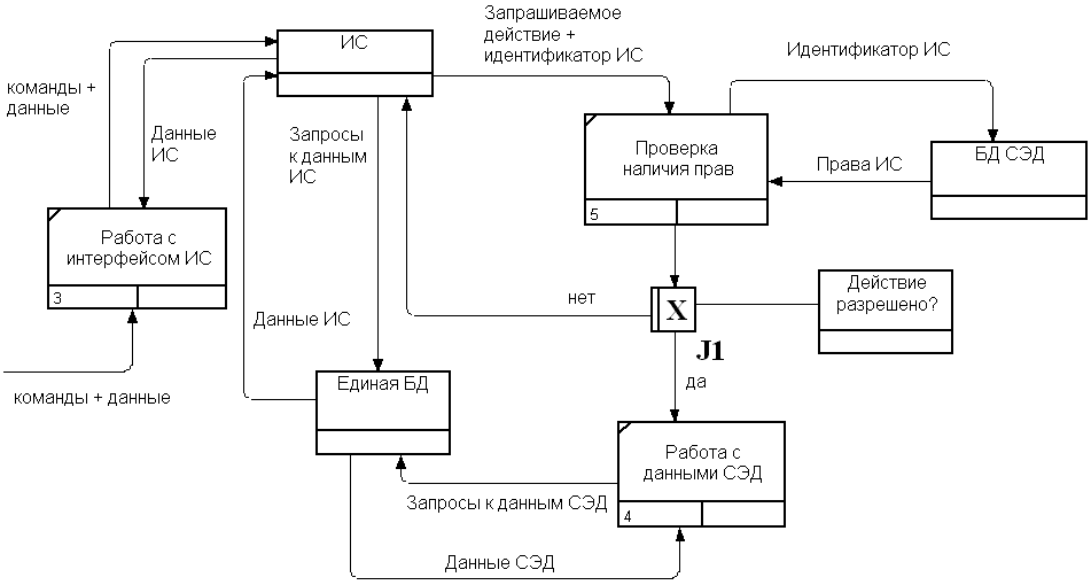

Рис. 3. Информационно-ориентированная интеграция

Для этого в СЭД разработан интерфейс, предоставляющий возможность управления данными СЭД в зависимости от назначенных другим ИС прав. Также в СЭД предусмотрен механизм, позволяющий использовать интерфейсы других ИС для доступа к их данным. Данная модель обеспечивается вводом в функциональную модель СЭД блоков импорта, экспорта и управления доступом.

В основу *модели сервисно-ориентированной интеграции* положено понятие ЭД, как объекта, способного содержать различную информацию как в структурированном, так и в слабоструктурированном виде (содержимое ЭД), снабжённого стандартизованным описанием (метаданными ЭД). Это позволяет использовать ЭД как контейнер данных произвольного типа и размера (см. Рис. 4). Кроме того, СЭД обладает набором сервисов, позволяющих каждой зарегистрированной в ней ИС работать с этими ЭД. Это обеспечивается вводом в функциональную модель СЭД блоков импорта, экспорта, управления доступом, работы с ЭД, а также выполнения дополнительных функций.

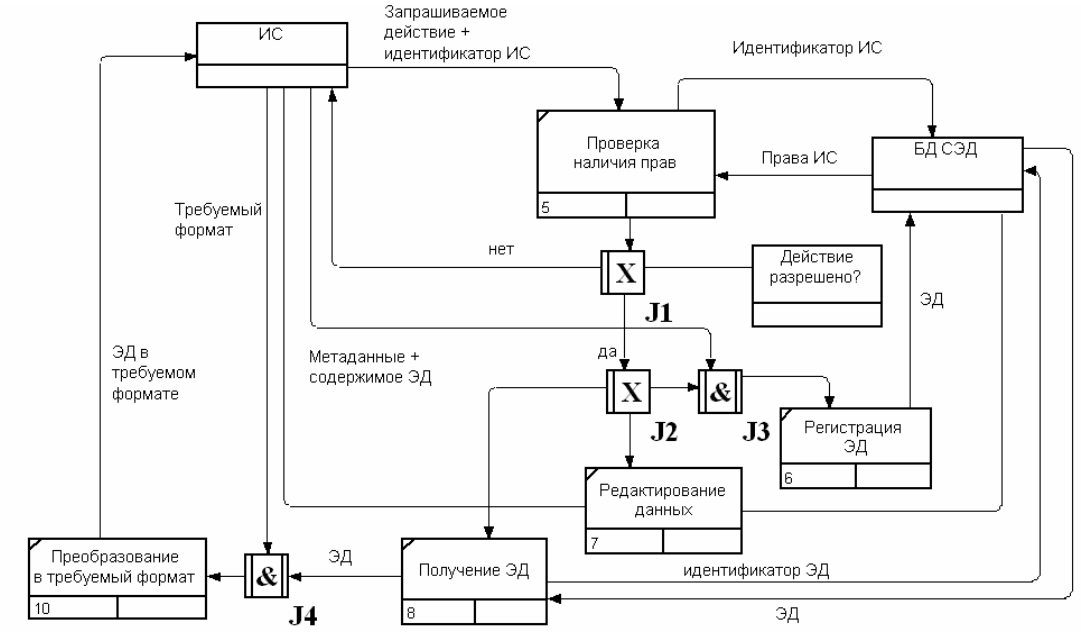

Рис.4. Сервисно-ориентированная интеграция

В основу *модели процессно-ориентированной интеграции* положено понятие БП, функциональных блоков управления и доступом к нему (см. Рис. 5).

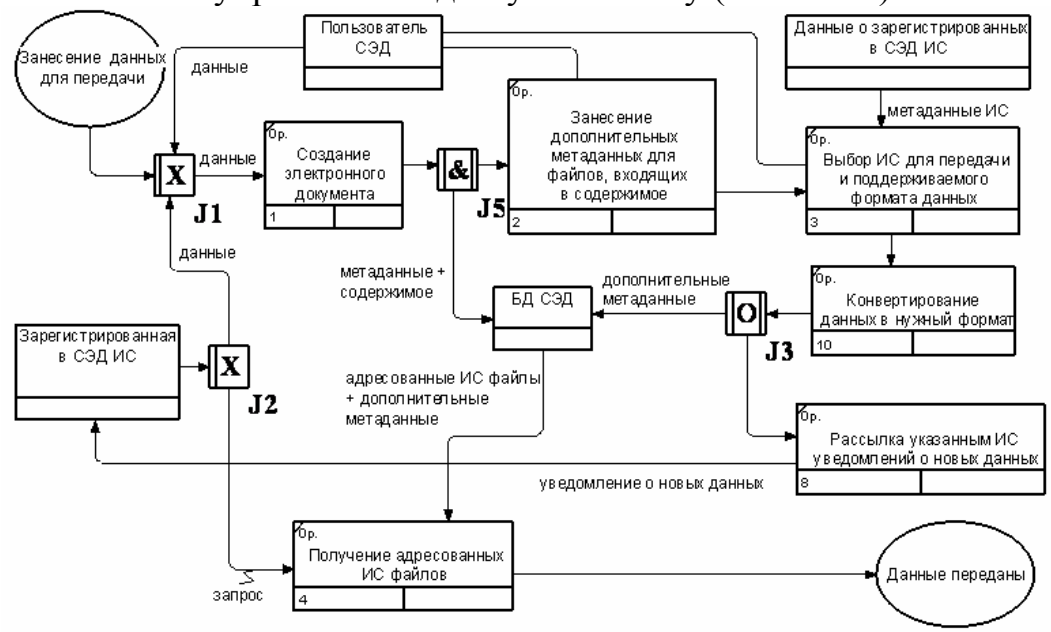

Рис. 5. Процессно-ориентированная интеграция

Интеграция осуществляется путём создания типовых БП, осуществляющих связь БП систем. Кроме того, предусмотрено создание индивидуальных БП, способных в ходе выполнения обеспечивать передачу данных в каждую ИС, либо получение данных из нее. Для эффективной реализации интеграции в СЭД разработан механизм создания БП на основе структуры их xml-описания.

Для удобства реализации процессно-ориентированной интеграции разработан язык KemSUBPDL, позволяющий формализовать элементы БП и организовать взаимодействие между ними, обладающий средствами поддержки вложенных процессов, предоставляющий возможности реализации ветвлений и циклов.

На Рис. 6 приведено RDF-представление интеграции ИС на основе СЭД.

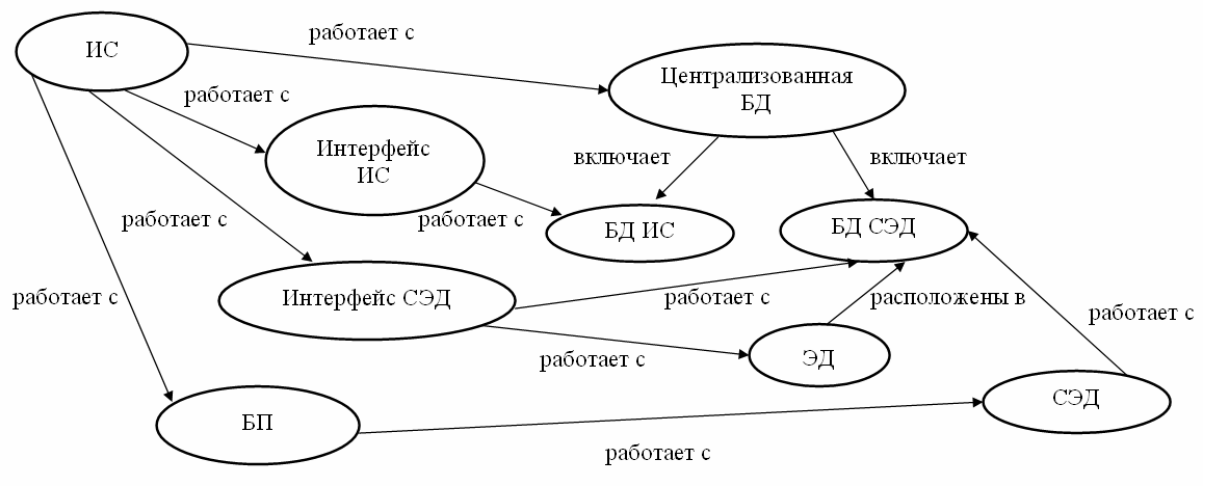

Рис. 6. Интеграция информационных систем

Поскольку СЭД оперирует в первую очередь с метаданными, их определение играет важную роль. В СЭД существует три метода определения метаданных: “ручной” (метаданные заносит сам пользователь), “автоматический” (метаданные автоматически определяются из содержимого ЭД) и “формальный” (метаданные определяются из описания, представленного в специально разработанном xml формате).

предложен метод автоматического определения метаданных, включающий несколько этапов:

1. **Выделение зон документа.** Под зоной документа понимается часть текста документа, имеющая единую структуру и семантическое значение.
2. **Определение типов зон документа.** Каждая зона имеет определённый тип, характеризующий её содержимое и множество метаданных, которые можно определить в зоне данного типа (например, заголовки, тело документа, списки, обращения).
3. **Определение класса документа.** Выполнение данного этапа основывается на понятиях матрицы и шаблона документа. *Матрица документа* представляется в виде $A = \begin{pmatrix} b_1 & Z_1 & l_1 \\ \dots & \dots & \dots \\ b_n & Z_n & l_n \end{pmatrix}$, где Z_i – тип *i*–ой зоны документа, l_i – длина *i*–ой зоны документа в символах, b_i – положение начала зоны относительно начала документа в символах. *Шаблоном документа* называется вектор $(Z_1,...,Z_n)$, определяющий порядок зон, где Z_i – тип *i*-ой зоны. Для одного класса ЭД может существовать множество шаблонов, описанных указанным методом. Класс ЭД определяется посредством сопоставление матрицы ЭД и набора возможных шаблонов.
4. **Последовательное определение метаданных в каждой зоне документа.** Основано на существовании для найденного класса шаблона, в котором указан набор зон и список возможных метаданных для каждой зоны.

На 2-4 этапах используется математический аппарат, основанный на построении нейронной сети Кохонена, применяемой для классификации образов путём нахождения близости параметров образа к ядру каждого класса. Для использования данного механизма определяются: ядра классов *C*, вектор входных параметров *x* и вектор *d*, характеризующий расстояние от *x* до *C*. Для определения класса, к которому относится объект, вы-

бирается минимальная евклидова норма d, т.е. решается задача $\|d(x,C)\| \to \min$. Соответствующий минимальному расстоянию класс является искомым.

На слой нейронов подается вектор входных параметров $x = (x_1, ..., x_n)$, где x_i – слова выбранной зоны документа, n – количество слов в зоне.

Ядро класса для метаданного имеет вид:

$$C = (k, r, \{s_1, ..., s_v\}, \{f_1, ..., f_u\}, \{t_1, ..., t_w\}) \quad (1)$$

где k – определяющий признак класса, r - позиция признака в тексте, $\{s_1, ..., s_v\}$ – ключевые слова, v – количество ключевых слов, $\{f_1, ..., f_u\}$ – ключевые фразы, u – количество ключевых фраз, $\{t_1, ..., t_v\}$ – шаблоны, соответствующий синтаксису регулярных выражений, w – количество шаблонов.

Задача поиска минимального расстояния $\|d(x,C)\|$ от вектора входных параметров до ядра класса сводится к определению:

$$q_1 d_1 + d_2 q_2 + d_3 q_3 + d_4 q_4 + d_5 q_5 \to \max, \quad (2)$$

где

- d_1 – величина, определяющая присутствие в документе определяющего признака, нормированное к числу слов зоны документа:

$$d_1 = \frac{1}{n} \sum_{i=1}^{n} [k = x_i]. \quad (3)$$

- d_2 – величина, определяющая расстояние между позицией определяющего признака в документе и позицией определяющего признака из ядра класса, нормированное к числу слов зоны документа:

$$d_2 = \frac{1}{n} \sum_{i=1}^{n} [k = x_i] \left(1 - \frac{|i - r|}{\max(i, r)} \right). \quad (4)$$

- d_3 – величина, определяющая количество слов зоны документа, совпавших со словами из ядра класса, нормированное к числу слов зоны документа:

$$d_3 = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{v} [s_k = x_i]. \quad (5)$$

- d_4 – величина, определяющая количество фраз зоны документа, совпавших с фразами из ядра класса, нормированное к числу слов зоны документа:

$$d_4 = \frac{1}{n} \sum_{i=1}^{n-m} \sum_{k=1}^{u} \left[f_k = \bigcup_{j=i}^{i+m} x_j \right]. \quad (6)$$

- d_5 – величина, определяющая количество слов, соответствующих регулярным выражениям из ядра класса, нормированное к числу слов зоны документа:

$$d_5 = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{w} \left[match(t_k, \bigcup_{j=i}^{p_k} x_j) \right], \quad (7)$$

где p_k – количество слов в *k-ом* регулярном выражении, *match (x,y)* – логическая функция, истинная, если фраза y соответствует регулярному выражению x и ложная в противном случае.

- весовые коэффициенты q_i удовлетворяют условию $\sum_{i=1}^{5} q_i = 1$.

Для эффективного определения метаданных в механизм включена функция обучения, состоящая в автоматическом добавлении новых ядер классов (автоматическое обучение), а так же в модернизации и добавлении шаблонов вручную (обучение с учителем).

Достоверность определения метаданных в период тестовой эксплуатации при занесении документов в СЭД составила 94% для структурированных (например, распоряжение, приказ, заявление) и 76% для слабоструктурированных (например, письмо, служебная записка) документов.

Метод выбора архитектуры СЭД основан на решении задачи оптимизации стоимости документопотоков, в основе которой лежит «архитектурный» подход – анализ возможного использования различных архитектурных решений при проектировании СЭД.

В качестве целевой функции выбрано выражение, включающее определение стоимостей хранения информации, получения данных из ИС и получения данных из БД СЭД при заданных ограничениях. Показано, что решение задачи оптимизации можно найти путём оценки эффективных параметров различных архитектурных решений, возможных для конкретного предприятия. Представлен анализ каждого из слагаемых целевой функции с учётом существенности полученных ограничений. На основе данного анализа разработан и представлен метод выбора оптимальной архитектуры СЭД.

На основе построенных моделей разработана структура данных СЭД, представленная в виде ER-модели.

Построенные модели соответствуют разработанным требованиям:

1. **Модели должны обеспечивать возможность эффективного добавления, редактирования, удаления и работы с данными произвольного типа и размера.** Данное требование обеспечивается за счёт ввода в информационную модель понятия ЭД, как объекта, способного содержать различную информацию как в структурированном, так и в слабоструктурированном виде, снабжённого стандартизованным описанием, а также вводом в функциональную модель блоков добавления, редактирования и удаления объектов содержимого.
2. **Модели должны обеспечивать возможность совместного с другими ИС использования одной и той же информации.** Данное требование обеспечивается вводом в информационную модель понятий ИС, а в функциональную модель блоков импорта, экспорта и управления доступом.
3. **Модели должны обеспечивать возможность интеграции ИС за счёт использования стандартизованного описания формата передачи данных и наличия у систем схожих сервисов работы с этим форматом.** Данное требование обеспечивается вводом в информационную модель понятия ЭД, отношений между ЭД, а в функциональную модель - блоков импорта, экспорта, управления доступом, выполнения дополнительных функций.
4. **Модели должны обеспечивать возможность объединения внутренних прикладных процессов каждой ИС за счёт создания в СЭД типовых бизнес-процессов.** Данное требование обеспечивается за счёт ввода в информационную модель понятий БП, процесса, задания и функции, а в функциональную модель блоков управления бизнес-процессами и управления доступом.
5. **Модели должны обеспечивать возможность автоматического определения метаданных из содержимого текстового документа, а также возможность определение**

метаданных из описания, представленного в специальном формате (например, в XML формате). Данное требование обеспечивается вводом в информационную модель понятий метаданных ЭД и XML формата их представления, вводом в функциональную модель блоков автоматического определения метаданных, в основу которого положено построение нейронной сети Кохонена, определения метаданных из XML формата.

6. **Модели должны обеспечивать возможность управления документами.** Данное требование обеспечивается вводом в информационную модель понятий ЭД, содержимого ЭД, метаданных ЭД, а в функциональную - модель блока работы с ЭД.
7. **Модели должны обеспечивать возможность маршрутизации.** Данное требование обеспечивается вводом в информационную модель понятия ЭД, маршрута ЭД, а в функциональную модель - блока управления БП. Маршрутизация ЭД осуществляется посредством создания связанного с ним процесса.
8. **Модели должны обеспечивать возможность управления бизнес-процессами.** Данное требование обеспечиваются вводом в информационную модель понятий БП, а в функциональную модель - блока управления БП.
9. **Модели должны обеспечивать возможность реализации удобных средств навигации и организации доступа пользователей к информации.** Данное требование обеспечивается за счёт ввода в информационную модель понятий папки, закладки, отношения, группы объектов, а в функциональную модель - блоков управления папками, закладками и блока осуществления различных видов поиска.
10. **Модели должны обеспечивать безопасность хранения и передачи информации.** Данное требование обеспечивается за счёт ввода в информационную модель понятий пользователя, группы пользователей, права и роли, а в функциональную модель блоков авторизации, разграничения прав доступа к объектам и делегирования прав.
11. **Модели должны предусматривать настройку СЭД на выбранную архитектуру.** Данное требование обеспечено предложенным методом выбора архитектуры СЭД на основе решения задачи оптимизации стоимости документопотоков.

Третья глава посвящена описанию реализации СЭД согласно предложенным моделям и методам.

С учётом построенных моделей СЭД, используя метод выбора оптимальной архитектуры системы, осуществлён выбор трёхзвенной архитектуры СЭД с централизованным управлением данными и доступом к системе через Интернет. В качестве клиентского приложения выбран максимально “тонкий” клиент - стандартный веб-броузер. Все электронные документы СЭД хранятся без дублирования в БД ЭД. Так как БД ЭД должна быть в некотором смысле “распределённой”, то есть состоять из двух частей – реляционной базы данных, предназначенной для хранения метаданных, и объектной базы данных, ориентированной на работу с документами произвольного типа, то в качестве БД была выбрана объектно-реляционная СУБД. Структура реляционных таблиц соответствует модели структуры данных. При реализации уровня бизнес-логики были использованы средства, обеспечивающие эффективную работу с данными, –пакеты PL/SQL и JAVA.

Для обеспечения доступа к логике СЭД из Интернета использован сервер приложений Apache Tomcat, выполняющий также функции web-сервера. Для реализации интерфейса СЭД использована библиотека KemSUWEB, разработанная в ЦНИТ КемГУ,

обеспечивающая единую среду для создания приложений, основанных на трехзвенной архитектуре, за счет адаптеров, которые удовлетворяют потребностям разработчика: в операциях с БД, в защите информации, в управлении ходом приложения. Интерфейс СЭД расположен на сервере приложений и представлен в виде набора xml-файлов, хранящихся в папках, имеющих древовидную структуру.

В конце главы приводится описание интерфейса и функционала реализованной СЭД, а также подтверждение соответствия созданной системы построенным моделям. Проведено тестирование СЭД на соответствие предъявленным к ней требованиям. СЭД внедрена в тестовую эксплуатацию в КемГУ и доступна по адресу http://sed.kemsu.ru.

В заключении приводятся основные результаты, полученные в диссертации, формулируются выводы, вытекающие из проведённых исследований, приводится ряд возможных направлений дальнейших исследований.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ, ПОЛУЧЕННЫЕ В ДИССЕРТАЦИИ

1. Предложена модель комплексной интеграции информационных систем, обеспечивающая информационно-ориентированный, сервисно-ориентрованный и процессно-ориентированный принципы интеграции.
2. Разработан перечень требований к моделям СЭД на основании результатов проведённого анализа существующих подходов к интеграции информационных систем и систем электронного документооборота. На его основе определены информационные объекты системы.
3. Разработаны информационная и функциональные модели, а также модели структуры данных СЭД, удовлетворяющие разработанным требованиям. В информационной модели описаны информационные объекты СЭД, а также отношения между ними. Набор функциональных моделей создан в виде совокупности диаграмм IDEF0 и IDEF3, отражающих функционал СЭД и существующие в системе потоки данных.
4. Разработан метод выбора архитектуры СЭД на основе решения задачи оптимизации стоимости документопотоков.
5. Предложен метод автоматического определения метаданных ЭД, в основу которого положена математическая модель нейронной сети Кохонена. Метод реализован и апробирован как часть СЭД.
6. На основе представленных моделей и методов реализована СЭД и проведена её апробация в процессе управления ВУЗом.

Достоверность результатов диссертационной работы подтверждается реализацией на основе построенных моделей в соответствии с разработанными требованиями СЭД и внедрении её в опытную эксплуатацию в КемГУ.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Журналы, рекомендованные ВАК для представления основных научных результатов диссертации:

1. Гудов, А. М. Об одной модели оптимизации документопотоков, реализуемой при создании системы электронного документооборота [Текст] / Гудов А. М., Завозкин С.Ю. // Вычислительные технологии. - 2006. - том 11, специальный выпуск. – С. 53 - 65

2. Гудов, А.М. Создание компонента автоматического определения метаданных документа для системы электронного документооборота [Текст] / А.М. Гудов, С.Ю. Завозкин, А.С. Меньшиков // Вестник томского государственного университета. – 2006. - №16.- С. 19-24
3. Гудов, А.М. Информационные и математические модели, заложенные в систему электронного документооборота КемГУ [Текст] / А. М. Гудов, С. Ю. Завозкин // Вестник НГУ.- 2005.- Том.2, вып.1, серия "Информационные технологии в образовании". – С. 68 – 73.
4. Завозкин, С. Ю. Процессно-ориентированная интеграция приложений при помощи системы электронного документооборота [Текст] / А. М. Гудов, С. Ю. Завозкин // Вестник ТГУ. – 2006. - № 19. - Приложение. Материалы международных, всероссийских и региональных научных конференций, симпозиумов, школ, проводимых в ТГУ. – С. 20 – 27.

Труды конференций:

1. Завозкин, С. Ю. Система электронного документооборота ВУЗа [Текст] / А. М. Гудов, С.Ю. Завозкин // Труды VI Всероссийской научно-практической конференции "системы автоматизации в образовании, науке и производстве". – Новокузнецк: СибГИУ, 2007. С. 278 – 281.
2. Гудов, А.М. Интеграция распределённых приложений при помощи системы электронного документооборота [Текст] / А.М. Гудов, С.Ю. Завозкин // Труды международной конференции "Вычислительные и информационные технологии в науке, технике и образовании". – Павлодар: ТОО НПФ "ЭКО", 2006. - II том. С. 442 – 451.
3. Завозкин, С.Ю. Об одном подходе построения архитектуры для реализуемой системы электронного документооборота ВУЗа [Текст] / С.Ю. Завозкин // Труды V Всероссийской научно-практической конференции "Недра Кузбасса. Инновации.". – Кемерово: ИНТ, 2006. С. 172 – 176.
4. Гудов, А. М. Создание моделей и методик для построения типовой системы электронного документооборота ВУЗа [Текст] / А. М. Гудов, С. Ю. Завозкин // Сборник материалов Всероссийского конкурса инновационных проектов аспирантов и студентов по приоритетному направлению развития науки и техники "информационно-телекоммуникационные системы". – М.: ГНИИ ИТТ "Информика", 2005. – С. 36-37
5. Гудов, А.М. Система электронного документооборота как элемент интеграции информационных систем ВУЗа [Текст] / А.М. Гудов, С.Ю. Завозкин // Труды V Всероссийской научно-практической конференции "Система автоматизации в образовании, науке и производстве". - Новокузнецк: СибГИУ, 2005. – С. 172 – 173.

Тезисы:

1. Завозкин, С.Ю. Анализ системы документопотоков ВУЗа с точки зрения эффективности передачи электронных документов [Текст] / С.Ю. Завозкин // Тезисы XI Международной научно-методической конференции "Новые информационные технологии в университетском образовании". - Кемерово: ИНТ, 2006. – С. 272 – 275.
2. Завозкин, С.Ю. СЭД как элемент интеграции подсистем ВУЗа в единую информационную систему [Текст] / С.Ю. Завозкин // Программа и тезисы докладов VI Всероссийской конференции молодых ученых по математическому моделированию и информационным технологиям (с участием иностранных ученых). - Кемерово, 2005. С. 62-63.