

Титов Ростислав Николаевич

**РАЗРАБОТКА СИСТЕМЫ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА
ДЛЯ АВТОМАТИЗАЦИИ УПРАВЛЕНИЯ НАУЧНО-ТЕХНИЧЕСКИМИ
ПРОЕКТАМИ (НА ПРИМЕРЕ ЕВРОПЕЙСКОЙ ОРГАНИЗАЦИИ
ЯДЕРНЫХ ИССЛЕДОВАНИЙ)**

Специальность 05.13.01 - системный анализ, управление
и обработка информации (в информационных системах)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва 2011

Работа выполнена в Европейской организации ядерных исследований и Национальном исследовательском ядерном университете «МИФИ».

Научный руководитель: доктор технических наук, профессор
Модяев Алексей Дмитриевич

Официальные оппоненты: доктор технических наук, профессор
Топорков Виктор Васильевич

кандидат физико-математических наук, с.н.с.
Кореньков Владимир Васильевич

Ведущая организация: Институт проблем информатики Российской академии наук (ИПИ РАН)

Защита состоится «29» февраля 2012 года в 16³⁰ часов на заседании диссертационного совета Д 212 130 03 в Национальном исследовательском ядерном университете «МИФИ» по адресу 115409, г. Москва, Каширское шоссе, дом 31.

С диссертацией можно ознакомиться в библиотеке Национального исследовательского ядерного университета «МИФИ».

Автореферат разослан: « ____ » _____ 2012 г.

*Ученый секретарь
диссертационного совета,
доктор технических наук, доцент*

Леонова

Н.М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность исследования. В настоящее время крупные научно-производственные объединения и организации, занимающиеся проектированием и изготовлением высокотехнологичных образцов новой техники, все в большей мере начинают применять в своей деятельности процессный подход. Это обусловлено, прежде всего, масштабом и распределенным характером выполняемых работ, сложностью административных видов деятельности, наличием большого числа привлекаемых смежников и квалифицированного персонала и многими другими факторами. Поэтому успешное и своевременное выполнение проектных работ, изготовление и выпуск продукции с требуемыми характеристиками с соблюдением заданного объема финансового и материально-технического обеспечения требует постоянного контроля, координации и увязки множества бизнес-процессов и технологических операций. При этом каждый из реализуемых бизнес-процессов требует составления, оформления и использования разнообразной документации на выполняемые работы. Сквозной характер процессного управления, реализующего все этапы деятельности организации, начиная с подготовки научно-технических проектов и завершая выпуском готовой продукции, требует наличия единого информационного пространства, в котором осуществляется документальный учет и сопровождение выполняемых работ. В силу ряда причин при реализации проектов часто возникает необходимость корректировки управления потоками работ, что требует оперативного изменения текущих документов.

Анализ существующих информационных систем, подходящих для нужд Европейской организации ядерных исследований (ЦЕРН), выявил отсутствие в них приложений, обеспечивающих автоматизированный электронный документооборот и работу с электронными подписями. В результате возникла необходимость разработки новой системы электронного документооборота (СЭД) для управления потоками работ - EDH (Electronic Document Handling), что и было положено в основу выполняемых диссертационных исследований и разработок. Поэтому данная диссертационная работа является актуальной.

Крупный вклад в теорию и практику создания и внедрения информационных систем для автоматизации документооборота и управления бизнес-процессами внесли видные отечественные и зарубежные специалисты, в числе которых В.А. Конявский, А.И. Агеев, М.Р. Когаловский, J. Ferguson, V. Van der Aalst, Th. Schwentick и другие. В их работах была сформирована основа для дальнейшего обобщения теоретических и практических результатов в области информационных технологий электронного документооборота. Вместе с тем, практика показывает, что создание и совершенствование систем электронного документооборота для автоматизации крупных предприятий и органов государственного управления требует проведения дополнительного комплекса исследовательских и проектных разработок.

В основу диссертационной работы положены материалы многолетних исследований и разработок по рассматриваемой тематике, выполненных автором в Европейской организации по ядерным исследованиям (ЦЕРН, Женева, Швейцария), начиная с 1995 года.

Цели и задачи исследования. Целью исследования является разработка новых принципов и методов построения системы электронного документооборота и создание программного комплекса для автоматизации управления научно-техническими проектами.

В ходе исследования решались следующие задачи:

1. Проведение анализа работ, посвященных решению задачи автоматизации документооборота, представления и маршрутизации электронных документов.
2. Обоснование и выбор математического аппарата для построения и исследования моделей электронных документов.
3. Разработка моделей электронных документов.
4. Разработка алгоритмов обработки электронных документов, поддерживающих автоматизацию управления потоками работ.
5. Разработка системы поиска электронных документов с возможностью семантического полнотекстового поиска.
6. Разработка программного комплекса для автоматизированного документооборота.

Объектом исследования являются методы и средства информационной поддержки управления потоками работ на основе системы электронного документооборота ЦЕРН.

Предметом исследования являются электронные документы и архитектура систем электронного документооборота.

Методы исследования. При составлении моделей электронных документов применяется математический аппарат типизированных деревьев и элементы теории множеств.

Научная новизна работы заключается в том, что в результате проведенных исследований:

1. Разработана оригинальная архитектура системы электронного документооборота, реализующая обработку электронных документов с декларативным описанием структуры документов на языке XML.
2. Разработана и обоснована формальная модель экземпляров электронных документов с выделением общих и уникальных реквизитов документов.
3. Предложена и обоснована обобщенная модель электронных документов и процессов маршрутизации при управлении потоками работ.
4. Разработаны и исследованы алгоритмы обработки наборов документов, реализующие расширенный набор запросов и механизмов разграничения доступа.
5. Разработаны процедуры выборки электронных документов, реализующие возможность семантического полнотекстового поиска.

Практическая значимость определяется тем, что:

1. Разработан и внедрен программный комплекс обработки электронных документов, обеспечивающий автоматизацию управления потоками работ и позволяющий прослеживать и корректировать движение документов в режиме реального времени.
2. Разработаны способы адаптации систем электронного документооборота и настройки под изменяющиеся условия реализации научно-технических проектов и характеристик выпускаемой продукции.
3. Созданы программные средства, обеспечивающие поддержку и сопровождение системы электронного документооборота в течение жизненного цикла.

Положения, выносимые на защиту

1. Архитектура системы электронного документооборота, реализующая обработку электронных документов с декларативным описанием структуры документов на языке XML.
2. Модели форм и экземпляров электронных документов на языке XML.
3. Алгоритмы обработки электронных документов, поддерживающих автоматизацию управления потоками работ.
4. Программный комплекс системы электронного документооборота, поддерживающий автоматизацию управления потоками работ.

Обоснованность и достоверность результатов работы обеспечиваются математическим обоснованием предлагаемых моделей, алгоритмов, а также адекватной интерпретацией полученных результатов.

Апробация и внедрение работы. Основные положения диссертационной работы были представлены на ежегодных научных сессиях МИФИ 2003, 2005, 2006, международных телекоммуникационных конференциях молодых ученых и студентов «Молодежь и наука» (Москва, 2002, 2008), XVII международной конференции «Математика. Компьютер. Образование» (Ижевск, 2010), 1-й международной школе ОИЯИ/ЦЕРН по Грид и административным информационным системам (Дубна, 2010) и внутренних семинарах ЦЕРН. По теме исследования опубликованы 19 работ (в том числе 6 публикаций в научных журналах, рекомендуемых ВАК). Результаты научной работы и методика автора были внедрены в ЦЕРН и использованы для автоматизации 61 бизнес-процесса ЦЕРН. Отдельные элементы методики были применены также в системах Объединенного института ядерных исследований (ОИЯИ, Дубна, Россия): «Электронный фотоархив», «Менеджер презентаций» и «Система управления персоналом НРТ». По всем объектам и проектам имеются акты внедрения.

Структура и объём работы. Диссертация состоит из введения, четырёх глав, заключения, списка использованной литературы (140 источников) и приложения, включающего акты о внедрении. Работа изложена на 156 страницах основного текста. Содержит 12 таблиц, 37 рисунков.

Содержание работы

Во введении обосновывается актуальность темы диссертационной работы и приводится ее краткая характеристика. Формулируются цель работы и задачи исследования и представляются основные положения, выносимые на защиту.

В первой главе дан обзор имеющейся литературы по теме исследования и рассмотрены научно-методологические и технологические аспекты исследуемой задачи. В качестве основы информационной поддержки процессного управления выбрана автоматизация бизнес-процессов с использованием электронного документооборота. Раскрыта существующая понятийная база электронного документооборота и проведен сравнительный анализ существующих систем.

Внедрение СЭД позволяет решить следующие наиболее часто встречающиеся задачи: обеспечение более эффективного управления за счет автоматического контроля выполнения, прозрачности деятельности организации на всех уровнях; поддержку системы контроля качества, соответствующей международным нормам; поддержку эффективного накопления, управления и доступа к информации и знаниям; обеспечение кадровой гибкости за счет большей формализации деятельно-

сти каждого сотрудника и возможности хранения всей предыстории его деятельности; протоколирование деятельности предприятия в целом (внутренние служебные расследования, анализ деятельности подразделений, выявление "горячих точек" в деятельности); оптимизацию бизнес-процессов и автоматизацию механизма их выполнения и контроля; исключение бумажных документов из внутреннего оборота предприятия и связанную с этим экономию ресурсов за счет сокращения издержек на управление потоками документов в организации; исключение необходимости или существенное упрощение и удешевление хранения бумажных документов за счет наличия оперативного электронного архива.

В главе приводится классификация: СЭД, ориентированные на бизнес-процессы, предназначенные для обеспечения полного жизненного цикла работы с документами (Documentum, FileNet Panagon, Hummingbird PC DOCS и др.), корпоративные СЭД, обеспечивающие инфраструктуру для создания документов, коллективной работы над ними и их публикации (Lotus Domino, Novell GroupWise, Open Text LiveLink, Keyfile, Oracle Context, iManage и др.), системы управления содержанием, системы управления информацией, обеспечивающие агрегирование информации и ее доставку через Internet, системы управления изображениями/образами, предназначенные для конвертации отсканированной с бумажных носителей информации в электронную форму и системы управления потоками работ, предназначенные для маршрутизации потоков работ в рамках корпоративных структурированных и неструктурированных бизнес-процессов. Системы управления потоками работ используются для повышения эффективности и степени контролируемости корпоративных бизнес-процессов.

Особый интерес для исследования представляют структурированные электронные документы, реквизиты которых представлены в формате, позволяющем использовать автоматизированную обработку содержащейся в них информации. Это позволяет осуществлять контроль правильности заполнения реквизитов на всех этапах жизни документа и делает возможным электронный обмен информацией между компьютерными системами различных организаций без участия человека (B2B – Business to business data exchange).

Проанализирована классификация реквизитов структурированных электронных документов, применяемых в системах управления потоками работ, подобных EDH. Набор реквизитов, входящих в электронный документ, варьируется для различных бизнес-процессов, поскольку документы предназначены для решения различных задач. Однако можно выделить набор реквизитов, входящих в любой документ независимо от типа поддерживаемого этим документом бизнес-процесса. Стандарт ISO 15836:2009 определяет «Дублинское ядро», состоящее из 15 реквизитов, общих для всех электронных документов (заголовки, создатель документа, тема, дата создания, тип документа и др.). Помимо реквизитов, общих для всех документов, часто бывает возможным выделить реквизиты, общие для некоторого подмножества электронных документов организации. Например, все финансовые документы ЦЕРН (заявка на заказ оборудования, платежное поручение, поручение о переводе денег между отделами и др.), содержат реквизиты «стоимость» и «код бюджетной классификации», а документы по работе с персоналом (апробация сотрудника, заявка на отпуск, оплата сверхурочных и т.п.) содержат общий реквизит «сотрудник». Некоторые документы могут относиться сразу к нескольким подмножествам. Электронные документы содержат также уникальные реквизиты, отсутствующие в документах других бизнес-процессов. Каждый рек-

визит имеет определенный тип данных («строка», «дата» и т.д.). Реквизиты документов могут представлять собой и сложные понятия, выражаемые сложными (комплексными) типами данных (например «сотрудник»). Множество типов данных, используемых в СЭД, конечно и определяется особенностями организации и областью применения СЭД (предметной областью).

Управление потоками работ в СЭД осуществляется при помощи систем маршрутизации, которые могут либо входить в состав СЭД (TIBCO, Lotus), либо являться самостоятельным продуктом (Oracle BPEL, Active Endpoints, Twister/Agila BPEL, PXE, и др.). Основными задачами системы маршрутизации в СЭД является сопровождение процесса утверждения документов в соответствии с правилами исполнения бизнес-процессов, сбор необходимых электронных подписей и передача утвержденных документов в корпоративные системы. Формализация правил маршрутизации обычно осуществляется при помощи высокоуровневых сетей Петри, а для описания параллельной маршрутизации можно использовать математический аппарат пи-исчисления.

Объектом исследования является СЭД EDN, решающая следующие основные задачи: создание электронных документов в интерактивном режиме, контроль правильности заполнения документов, разграничение прав доступа, маршрутизацию документов в соответствии с требованиями административных бизнес-процессов, сбор электронных подписей, уведомление пользователей о движении документа, снижение задержек при утверждении путем перенаправления документов в случае отсутствия или бездействия должностных лиц, делегацию прав подписи, возможность прослеживать и корректировать движение документа в режиме реального времени, поиск документов, контроль бюджета в реальном режиме времени, автоматизированную передачу данных из документов в корпоративные системы ЦЕРН и фирм-поставщиков.

Рассмотрены проблемы, возникающие при решении задачи автоматизации бизнес-процессов в крупных организациях и НПО, подобных ЦЕРН. Показано, что можно создать общую модель, применимую к любому электронному документу вне зависимости от характера бизнес-процесса. Такая модель позволяет выделить ряд операций, присущих любому документу и создать подсистему обработки электронных документов, существенно ускоряющую автоматизацию новых бизнес-процессов и снижающих стоимость технического сопровождения системы. Наличие общей модели документов позволяет также применить новые подходы к поиску документов, с учетом их семантики и возможностей полнотекстового поиска. Суть предлагаемой методики заключается в замене разрозненных алгоритмов подсистемой обработки, упрощающей описание структуры и логики работы документов. Конечной целью метода является сокращение времени на разработку новых типов электронных документов, а, следовательно, и на автоматизацию новых бизнес-процессов в целом, при одновременном снижении стоимости технического обслуживания. Для моделирования электронных документов предлагается использовать язык XML, что, в частности, дает возможность использовать для поиска специализированные поисковые системы.

Рассматриваются основные критерии оценки качества использования электронного документооборота, где под качеством в первую очередь понимается способность СЭД удовлетворять нуждам пользователей. Основными стандартами качества ПО в настоящее время являются стандарты ГОСТ Р ИСО/МЭК 9126, определяющий номенклатуру, атрибуты и метрики требований качества, и новый стан-

дарт ISO/IEC 25000, определяющий модель качества, известную под названием SQuaRE (Software Quality Requirements and Evaluation – Требования к качеству ПО и его оценка). Поскольку качество является относительным понятием, для его количественной оценки необходимы метрики (количественные показатели). Ими могут быть конкретные значения, диапазоны изменения значений и модели их оценивания. Специальной мерой может служить уровень использования повторных компонентов, измеряемый как отношение размера продукта, изготовленного из готовых компонентов, к размеру системы в целом.

Простейшей метрикой размера ПО является количество строк исходного кода (SLOC – Source Lines of Code). Для сравнения метрик SLOC разных языков программирования обычно вводятся коэффициенты пересчета, например, в количество строк ассемблерного кода (KAELOC) или в количество строк некоторого абстрактного языка программирования. Эксперименты подтверждают хорошую корреляцию метрики SLOC с трудозатратами, что оправдывает использование этой метрики для оценки трудозатрат, хотя такая оценка дает адекватные результаты только при оценке производительности достаточно большой команды разработчиков с разным профессиональным уровнем. Другими размерно-ориентированными метриками являются метрики Холстеда, в основе которых лежат четыре измеряемые характеристики программы:

- $N_{U_{Opt}}$ – число уникальных операторов (словарь операторов);
- $N_{U_{Opd}}$ – число уникальных операндов (словарь операндов);
- N_{Opt} – общее число операторов;
- N_{Opd} – общее число операндов;

На основании этих характеристик рассчитываются следующие оценки:

- $P_{Voc} = N_{U_{Opt}} + N_{U_{Opd}}$ – словарь программы;
- $P_{Len} = N_{Opt} + N_{Opd}$ – длина программы;
- $P_{Vol} = P_{Len} \times \log_2 P_{Voc}$ – объем программы;
- $P_D = \frac{N_{U_{Opt}}}{2} \times \frac{N_{Opd}}{N_{U_{Opd}}}$ – сложность программы;
- $E = P_D \times P_{Vol}$ – оценка усилий программиста при разработке;

Одной из распространенных оценок сложности программных проектов является показатель цикломатической сложности Мак-Кейба, для вычисления которого необходимо построить ориентированный граф управляющей логики программы, где вычислительные операторы или выражения представляются в виде узлов, а передача управления между ними – в виде дуг. Обычно метрику Мак-Кейба используют для оценки сложности отдельных процедур. Цикломатическое число Мак-Кейба показывает требуемое количество проходов для покрытия всех контуров графа или количество тестовых прогонов программы, необходимых для исчерпывающего тестирования по принципу «работает каждая ветвь».

В зависимости от решаемой задачи, для оценки эффективности автоматизации документооборота могут быть применены различные показатели из числа рассмотренных выше.

Вторая глава посвящена разработке математической модели электронных документов. Документ, со своими реквизитами и группами реквизитов, по сути,

представляет собой иерархическую структуру данных, причем реквизиты комплексных типов данных в свою очередь могут быть представлены в виде иерархической структуры. К примеру, документ ЦЕРН «заявка на заказ оборудования» содержит секцию общих реквизитов, секцию параметров доставки (адрес доставки и др.) и список заказываемых позиций, каждая из которых в свою очередь может также представлять собой иерархическую структуру (рис. 1). Для формального описания иерархических структур данных существуют различные языки. В диссертации для моделирования электронных документов предлагается использовать широко распространенный язык XML, позволяющий структурировать информацию в виде иерархических деревьев при помощи специальных меток («тэгов»). В этом случае для обработки документов можно использовать технический аппарат XML: определение структуры данных при помощи XML-схем, преобразования XSLT, навигацию по дереву при помощи языка XPath и язык запросов XQuery.

Material Request 4521247 - Windows Internet Explorer

https://edh.cern.ch/Document/SupplyChain/MAG

Material Request 4521247

Заказ оборудования 4521247

Клон Прикрепить Печать Сроки? Сохранить Послать Помощь

Конфиденциально Срочно

Создатель документа: **Ростислав Титов (GS-AIS-EB)** тел. 79605, дата создания 01.12.2010

Поля, помеченные звездочкой (*) обязательны для заполнения

Общая информация

Описание *: Новый компьютер ?

Заказчик *: Ростислав Титов (GS-AIS-EB) ?

Код затрат *: 71101 ?

Параметры доставки

Доставить в офис *: 513-1-002 ?

Желаемая дата доставки: ?

Комментарий: ?

Контактная информация: cern.stores@cern.ch

Итого: 789.00 фр.

Доступные электронные каталоги (двойной щелчок для входа в каталог):

CERN Stores Farnell Radiospares Bossard SFS unimarket Lyreco Distrelec

Заказываемое оборудование

Поз.	Количество	Цена за	Каталожный №	Комментарий	Цена
1	1	=1.00 шт.	80.01.01.013.3	Компьютер DELL Optiplex 780 Standard desktop computer DELL Optiplex780-4GB RAM-2X2GB	733.00 X
2	1	=1.00 шт.	80.02.05.002.0	USB-Клавиатура фирмы HP USB HP KEYBOARD US layout	22.00 X
3	2	=2.00 шт.	80.02.05.050.6	Оптическая мышь для ПК/Мас OPTICAL ROLLER MOUSE for PC/MAC	34.00 X

Нажмите сюда для добавления новой позиции +Добавить

Рис. 1. Пример электронного документа ЦЕРН «Заявка на заказ оборудования»

Для моделирования документов XML использованы теория и математический аппарат типизированных деревьев.

Введем множество базовых типов $B = \{b_1, b_2, \dots, b_m\}$. Применительно к реальным документам, это могут быть как простые («текст», «число»), так и комплексные («сотрудник», «код затрат») типы данных.

Для каждого из элементов множества B определим область допустимых значений данного типа, $M(b_i)$.

Определение « Σ -дерево». Зафиксируем конечный алфавит Σ имен элементов. Множество Σ -деревьев (обозначаемое T_Σ), определяется индуктивно следующим образом:

- (1) каждый элемент $t \in \bigcup_{i=1}^m M(b_i)$ есть элементарное Σ -дерево;
- (2) если $\sigma \in \Sigma$ и $t_1, \dots, t_n \in T_\Sigma$, где $n \geq 1$, то $t = \sigma(t_1, \dots, t_n)$ есть Σ -дерево.

Здесь запись $\sigma(t_1, \dots, t_n)$ означает, что элемент σ является родительским по отношению к каждому из элементов $t_1, \dots, t_n$, сами же эти элементы расположены на одном уровне («элементы-братья», siblings), поскольку имеют общего родителя.

Стоит отметить, что в вышеприведенном определении количество дочерних элементов для данной вершины никак не ограничено. В дальнейшем в этом разделе для краткости будем под понятием «дерево» понимать « Σ -дерево».

Рассмотрим концепцию типизации узлов дерева. Будучи иерархической структурой, дерево XML включает в себя поддеревья, при этом узлы дерева с одинаковыми именами (например, «сотрудник») могут встречаться на разных уровнях дерева. Типизация узлов дерева позволяет различать такие узлы и, при необходимости, обрабатывать эти узлы по-разному (например, в задачах контекстного поиска по дереву).

Любая листовая вершина документа σ может иметь один из базовых типов, принадлежащих множеству B :

$$\exists \tau \in B : \sigma \in M(\tau),$$

В терминах типизированных деревьев данное выражение может быть также описано следующим образом (читается «дерево σ имеет тип τ »):

$$\exists \tau \in B : (\sigma : \tau).$$

Для определения типов элементов в XML обычно используются XML-схемы или определения типа документа (DTD – Document Type Definition). Обозначим как E множество типов элементов, описанных в схеме, где каждому элементу соответствует определенное имя.

Пусть $n_1, n_2, \dots, n_k$ – это путь в дереве от корневой вершины n_1 к листовой вершине n_k , причем $\forall i: i \in (1; k-1), n_i$ – родительский узел для n_{i+1} .

Тогда любой частичный путь $n_i, n_{i+1}, \dots, n_{i+j}$, где $i \in (1; k-1), j \geq 0, j \leq k - i$, представляет собой подтип.

Полное множество подтипов E' – это (возможно бесконечное) множество типов элементов и их подтипов. В отличие от множества E , содержащего только типы с простыми именами, множество E' будет содержать также подтипы, имеющие составные имена (то есть $E' \subseteq E$).

Рассмотрим способы моделирования электронных документов при помощи типизированных деревьев. Введенное выше понятие Σ -дерева можно использовать для описания экземпляра электронного документа. Листовые вершины дерева, соответствующие случаю (1) в определении, представляют собой вершины, в которых хранятся данные реквизитов документа, а вершины, соответствующие случаю (2) можно рассматривать как контейнеры, которые не содержат данных и служат для логической группировки реквизитов.

В связи с необходимостью регламентировать структуру экземпляров документов обратимся к понятию типа дерева и элемента дерева. Стоит отметить, что один и тот же элемент может быть связан с несколькими типами из различных

множеств: узел дерева документа может иметь элементный тип из множества E , соответствующий имени элемента (например, «имя»), и с другой стороны, может иметь несколько подтипов из множества E' (например, «создатель/имя», «заказчик/имя» и т.д.).

Во введенной системе обозначений графическое представление экземпляра документа рис. 1 может выглядеть так, как показано на рис. 2.

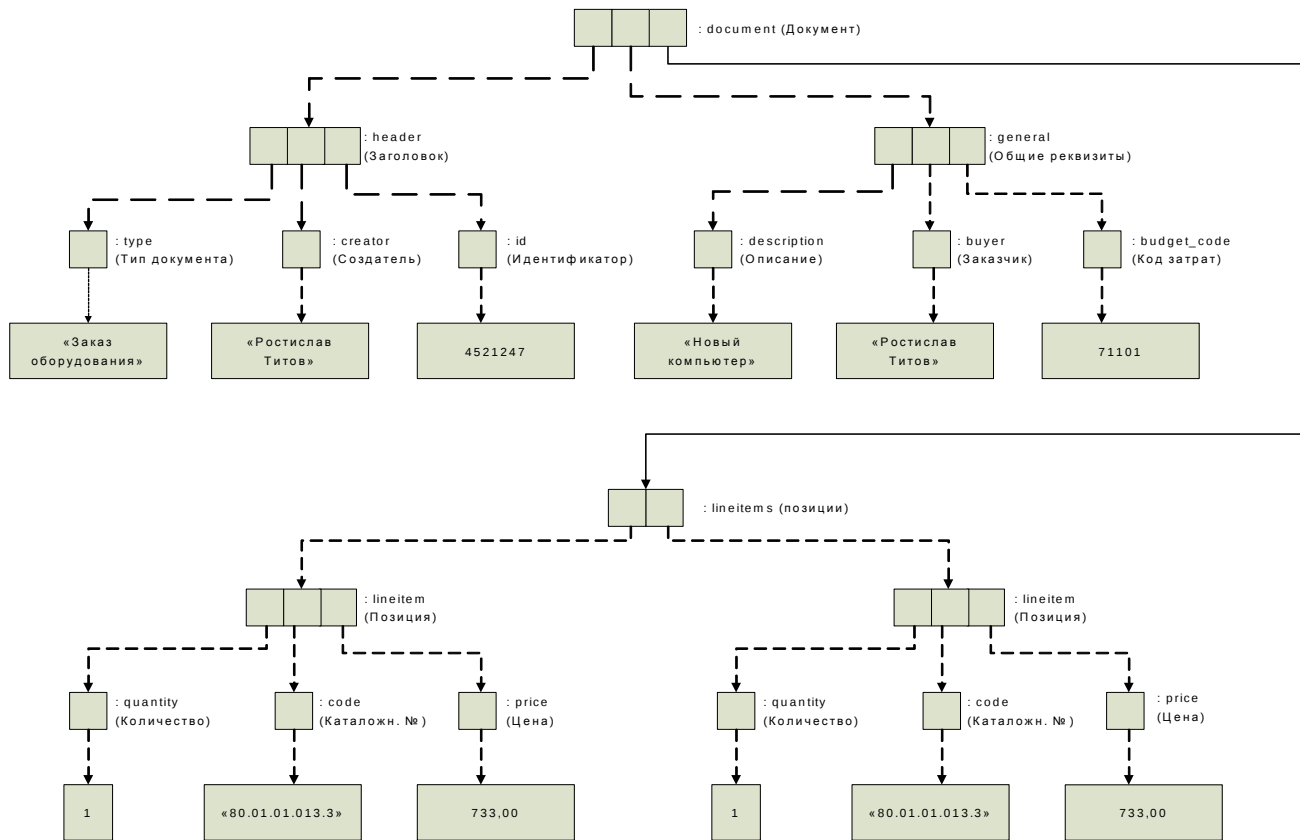


Рис. 2. Графическое представление экземпляра документа в виде типизированного дерева

Для решения практических задач поиска по дереву в диссертации вводятся отношения содержания и вхождения. Отношение содержания является транзитивным, антисимметричным и рефлексивным, а также не обладает свойством полноты. Это отношение нестрогого частичного порядка:

транзитивность: $\forall t_1 : \tau_1, t_2 : \tau_2, t_3 : \tau_3 \quad (t_1 \subseteq t_2) \wedge (t_2 \subseteq t_3) \Rightarrow t_1 \subseteq t_3$,

антисимметричность: $\forall t_1 : \tau_1, t_2 : \tau_2 \quad (t_1 \subseteq t_2) \wedge (t_2 \subseteq t_1) \Rightarrow t_1 = t_2$,

рефлексивность: $\forall t : \tau \quad t \subseteq t$.

Что касается отношения вхождения, то при его определении будем считать, что для всех базовых типов существует алгоритм определения вхождения слова в элемент данного типа. В случае с языком XML данная задача сводится к проверке присутствия последовательности символов слова в самом элементе.

Понятия типизированного дерева и поддеревя, а также оператор содержания предоставляют фундамент для описания алгебры поддеревьев, за основу которой берут реляционную алгебру. В качестве носителя алгебры поддеревьев выбирают множество наборов данных, поскольку операции проекции и выбора, вообще говоря, незамкнуты относительно множества баз структурированных данных (так

как результат такой операции не всегда включает в себя корень дерева). Алгебра на поддеревьях определена как пара:

$$A = \langle D, S \rangle ,$$

где A – алгебра поддеревьев,

D – носитель алгебры (множество наборов данных),

S – сигнатура алгебры – множество операций над элементами носителя, содержащее нижеприведенные операции.

Пусть D – это набор данных. Операцию проекции (обозначаемую $\rho_T(D)$, где T – множество типов) определяют следующим образом:

$$\rho_T(D) = \{t : t \in D \mid t \in T\} .$$

Проекция произвольной глубины (обозначаемая $\rho_T^*(D)$), может быть определена как:

$$\rho_T^*(D) = \{\rho_T(D) \cup \{t : t \in D \mid t \subseteq \rho_T(D)\}\}$$

Пусть D – это набор данных. Операцию выбора (обозначаемую $\sigma_{t'R}(D)$, где t' – поддерево, R – бинарный булев оператор) определяют следующим образом:

$$\sigma_{t'R}(D) = \{t : t \in D \mid t' R t\}$$

В качестве отношения R можно использовать одно из определенных выше отношений: содержания ($\subseteq$) и вхождения ($\subset$). Так, например, запрос:

$$\sigma_{\text{"document"} \subseteq} (D)$$

выберет все деревья из D , включающие слово «document» на любом уровне глубины.

Вводимые операции позволяют сформулировать широкий набор запросов к документу. На практике при построении запросов, как правило, используется другой синтаксис, однако этот синтаксис должен поддерживать базовые операции алгебры поддеревьев. В частности, в данном проекте в качестве языка запросов используется XPath. Основные операции алгебры поддеревьев могут быть записаны и в синтаксисе XPath.

Операция проекции в XPath будет выглядеть, например, следующим образом:

//document/common

Данный запрос выберет все поддеревья, соответствующие подтипу «document/common», то есть будут выбраны все узлы с именем «common», дочерние по отношению к узлу «document», на каком бы уровне вложенности этот узел ни находился. В данном случае тип, по которому производилась проекция, являлся составным (из множества E'). Однако и проекция по простому элементному типу (из множества E) формулируется аналогично:

//common

Данный запрос выберет все узлы с именем «common», независимо от их положения в дереве документа. В обоих случаях в неявной форме используется ось child, являющаяся основной осью по умолчанию в языке XPath.

Проекция произвольной глубины, подразумевающая отбор не только деревьев заданного типа, но и всех их поддеревьев, легко формулируется в синтаксисе XPath с использованием оси descendant-or-self:

//common/descendant-or-self::*

Этот запрос выберет не только узлы с именем «common», но и все узлы, лежащие «ниже» по иерархии. Здесь знак * означает, что имя элемента-потомка может быть любым.

Операция выбора, аналогичная вышеприведенному примеру, может быть записана в XPath следующим образом:

```
//*[contains(., "document")]
```

Здесь знак * означает, что имя элемента может быть любым, а функция contains() применяется к контекстному узлу (который в синтаксисе XPath принято обозначать точкой) и действует аналогично отношению содержания.

В третьей главе рассматривается применение математического аппарата теории типизированных деревьев при автоматизации документооборота.

Для описания электронных документов удобно использовать две модели: модель структуры и модель представления. Это связано с тем, что документы и их представление в пользовательском интерфейсе системы имеют различные жизненные циклы: документ может быть создан до того, как пользователь получает к нему доступ (например, документ может быть сгенерирован автоматически другой компьютерной системой), и продолжает свою жизнь после отключения пользователей от системы. Декомпозиция модели документа на модель структуры и модель представления позволяет отделить модель данных от пользовательского интерфейса в соответствии с концепцией трехуровневой архитектуры и шаблоном проектирования «модель-представление-поведение».

Модель структуры документа предлагается использовать для описания структуры документа (реквизитов документа и типов хранимых в них данных) и прав доступа. Модель представления определяет, в каком виде электронный документ будет показан пользователю.

Для моделирования электронных документов используем деревья XML, а для описания внутренней логики, правил контроля доступа и взаимосвязей между элементами модели язык XPath, предназначенный для навигации по XML.

Приведем графическое представление модели структуры документа в виде XML-схемы (рис. 3).

Реквизиты документа описываются в элементах модели *field*, где можно указать также тип данных (*type*) и максимально допустимый размер (*size*) поля реквизита. Каждому реквизиту присваивается уникальное имя (*name*), используемое для его дальнейшей идентификации. Для списочных типов данных в структуре документа может быть задан набор допустимых значений. Множество допустимых типов данных зависит от области применения СЭД и задается отдельно.

Некоторые группы реквизитов («позиции») могут повторяться в электронном документе более одного раза (рис. 1), для чего в модель введен элемент *lineitems*, позволяющий задать неограниченное количество позиций и определить набор реквизитов, входящий в каждую из них. При этом можно задать максимальное и минимальное количество повторений для каждой позиции (например, в техническом задании может быть указано, что некоторые группы данных должны присутствовать в документе не менее одного и не более трех раз).

Для каждого реквизита электронного документа могут быть заданы права доступа при помощи элемента модели *constraints*. В разработанной модели реализовано три типа ограничений: READONLY – запрет на модификацию, MANDATORY – запрет на ввод пустых значений и ERROR – запрет на ввод значений, определяемых внутренней логикой. Для ограничений типа ERROR (ошибка)

может быть дополнительно задан код ошибки для представления в пользовательском интерфейсе. Ограничения задаются в виде неравенств на языке XPath и могут выражать зависимость от других полей документа, зависимость от нового значения текущего поля или любую другую зависимость (через вызов внешних функций языка XPath).

Аналогичным образом ограничения могут быть заданы для всего документа в целом, для чего в модели присутствуют элементы *read-access* для определения множества лиц, имеющих доступ к документу, и *create-access* для определения множества лиц, имеющих право создавать новые экземпляры документов.

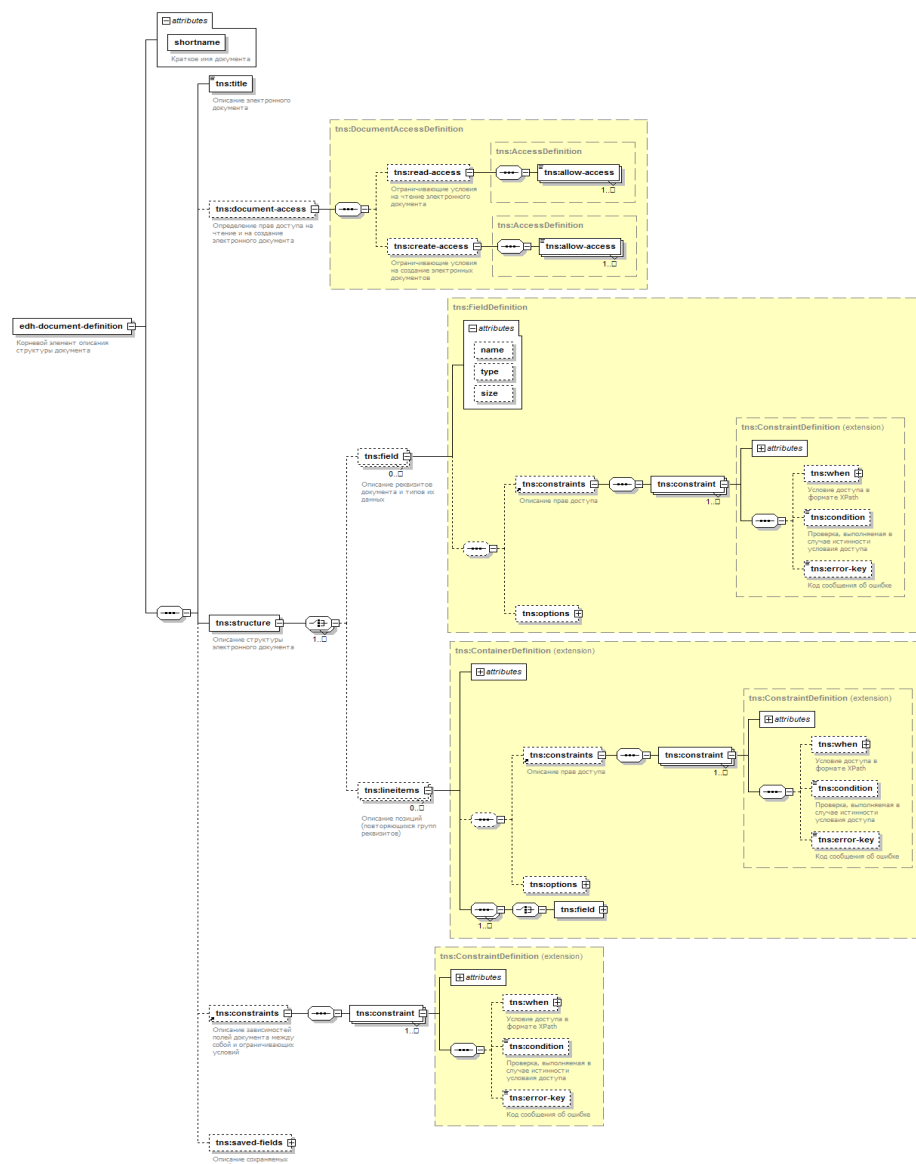


Рис. 3. Модель структуры документа

Модель представления документа используется совместно с моделью структуры для представления документа пользователю.

Как и модель структуры документа, модель представления содержит элемент описания прав доступа *constraints*, в котором, в дополнение к перечисленным выше ограничениям, можно использовать ограничение **HIDDEN** (поле невидимо для пользователя). Это позволяет в случае необходимости скрывать часть реквизитов электронного документа в процессе его маршрутизации (например, в ЦЕРН большинство должностных лиц, подписывающих платежную ведомость, не видят

банковских реквизитов получателя денег, в то время как сотрудники финансового отдела имеют к ним доступ).

Для моделирования экземпляров электронных документов, то есть заполненных документов, имеющих уникальный идентификатор и содержащих данные, используем Σ -деревья. В листовых вершинах дерева будем размещать данные реквизитов документа, а не-листовые вершины, соответствующие случаю (2) определения Σ -дерева, будем применять для логической группировки реквизитов. Предлагается выделить четыре вершины первого уровня для группировки общих реквизитов, реквизитов, общих для некоторых классов документов, уникальных реквизитов и информации о процессе маршрутизации. Поскольку набор уникальных реквизитов различен в зависимости от типа бизнес-процесса, внутренняя структура соответствующей вершины может варьироваться, что, как показано в диссертационной работе, может быть практически реализовано путем добавления дополнительной XML-схемы. С учетом вышесказанного, обобщенная модель экземпляра электронного документа может выглядеть так, как это показано (в виде графического представления XML-схемы) на рис. 4.

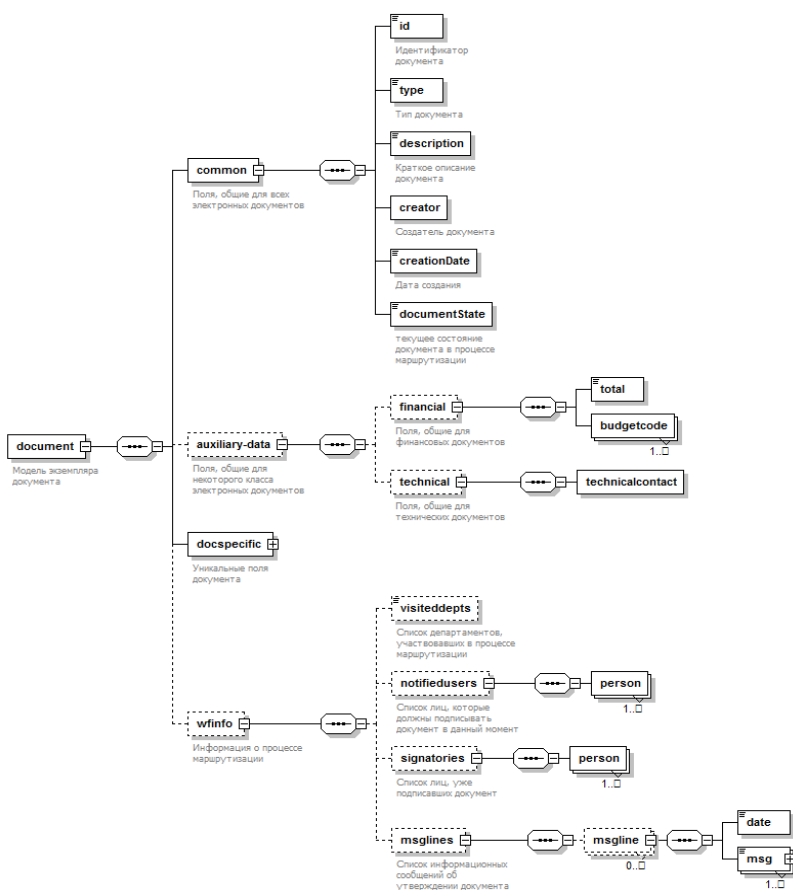


Рис. 4. Модель экземпляра электронного документа

Предлагаемая модель обладает следующими свойствами:

- может использоваться для представления экземпляров любых документов; для новых бизнес-процессов достаточно описать в модели структуру вершины для хранения уникальных полей.
- единообразное задание общих реквизитов устраняет несогласованность структур данных, описывающих различные электронные документы, что облегчает и упрощает хранение и поиск документов.

- модель не зависит от типов данных, что позволяет применять ее для представления экземпляров документов в различных СЭД.
- посредством задания XML-схемы для модели экземпляра и для используемых в модели типов данных, можно обеспечить проверку правильности формирования электронного документа.

Для сохранения комплексных типов данных в системе управления базами данных (СУБД) возможны три подхода: сохранение минимального набора атрибутов (в дерево экземпляра включаются только составляющие первичного ключа, уникально идентифицирующего сохраняемый объект); сохранение полного набора атрибутов (в дерево включаются все атрибуты); сохранение частичного набора атрибутов (сочетание двух предыдущих подходов).

Сохранение части атрибутов комплексных типов данных в модели экземпляра документа упрощает поиск информации. Анализ использования СЭД в ЦЕРН показывает, что для поиска документов чаще всего используется весьма ограниченный набор атрибутов, что обуславливает целесообразность сохранения частичного набора атрибутов при записи экземпляров документов в СУБД. Решение о том, какие атрибуты должны сохраняться для каждого комплексного типа данных, должно приниматься на основании экспертных знаний о предметной области и потребностях по поиску документов.

Применение математического аппарата типизированных деревьев позволяет перейти к иерархическим моделям документов и их экземпляров. В диссертации рассмотрено использование этих моделей для разработки модели подсистемы обработки электронных документов (ПОЭД). При проектировании этой подсистемы были применены архитектура «модель-представление-поведение» (MVC – Model-View-Controller) и принципы объектно-ориентированного программирования (ООП). ПОЭД реализована на языке Java с использованием сервлетного подхода. *Уровень модели* подсистемы содержит классы для обработки моделей электронных документов и их экземпляров. На этом уровне реализована функциональность, необходимая для построения модели экземпляра на основе XML-модели документа, проверки прав доступа, обработки бизнес-логики и взаимодействия с СУБД. *Уровень поведения («контроллер»)* представлен иерархией сервлетов и исполнительных объектов (ИО). ИО предназначены для сохранения внутреннего состояния электронных документов между запросами, поскольку сервлеты не предоставляют такой возможности. *Уровень представления* отвечает за представление электронных документов пользователю. Внешний вид электронных документов задается при помощи серверных страниц Java (Java Server Pages). Графическая диаграмма иерархии классов ПОЭД на языке UML приведена на рис. 5.

Разработанная модель ПОЭД обеспечивает не только горизонтальное разбиение функциональности по набору решаемых задач в соответствии с архитектурой «модель-представление-поведение», но и вертикальное разделение подсистемы на три подуровня по степени обобщения функций, расширяющих круг решаемых подсистемой задач. *Базовый подуровень* ПОЭД включает в себя функциональность, общую для всех экранов пользовательского интерфейса СЭД. *Подуровень электронных документов* реализует функции, общие для электронных документов различных типов, расширяя функциональность базового подуровня операциями обработки команд, специфичных для электронных документов, обработки электронно-цифровых подписей и контроля прав доступа на основе модели роле-

вого доступа RBAC. Подуровень обработки метаданных позволяет создавать новые электронные документы непосредственно на основе их моделей, заданных в виде XML-деревьев.

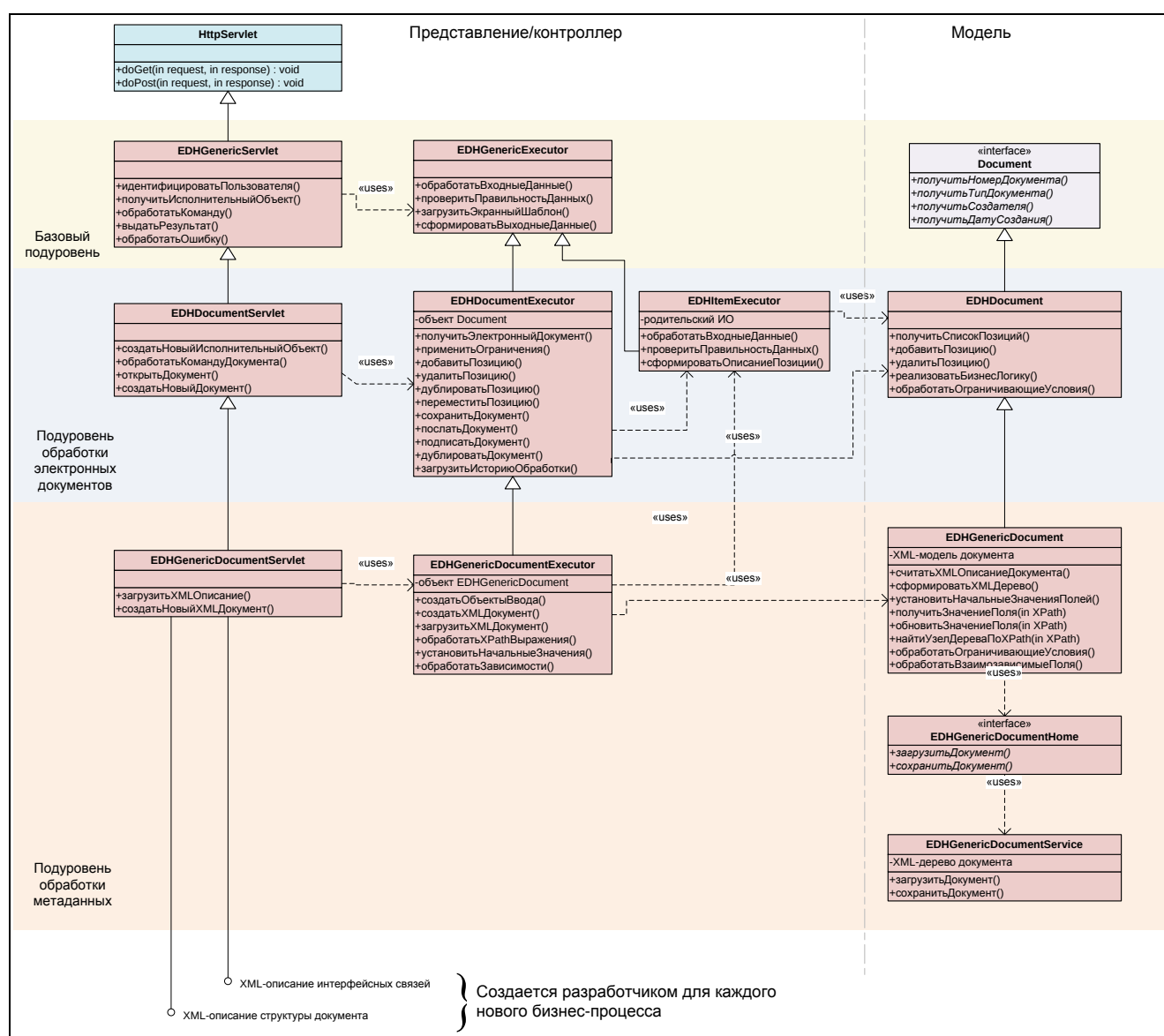


Рис. 5. Упрощенная UML-модель подсистемы обработки электронных документов. Класс `HttpServlet` входит в состав интерфейса сервлетов J2EE. `Document` является базовым интерфейсом СЭД EDH. Остальные приведенные на диаграмме классы разработаны автором в ходе настоящего исследования.

Четвертая глава диссертационной работы посвящена вопросам практической реализации подсистемы обработки электронных документов для автоматизации бизнес процессов. Здесь подробно рассмотрены способы сохранения электронных документов в СУБД, обработки XML-деревьев и представления документов в пользовательском интерфейсе, методы использования разработанных моделей в ПОЭД и для поиска документов, процесс выбора подходящего аппаратно-программного обеспечения для СЭД, обеспечивающего требуемые масштабируемость и переносимость, и проведен анализ эффективности предлагаемых решений.

В настоящее время СЭД работает на кластере из 6 персональных компьютеров фирмы HP под управлением Linux. Для передачи информации используются сетевые контроллеры Fast Ethernet с пропускной способностью 100 Мб/с.

Одним из основных системных компонентов СЭД является СУБД. Существуют различные подходы к сохранению объектов в базах данных: применение объектно-ориентированных системы управления базами данных (ООСУБД), объектно-реляционных систем управления базами данных (ОРСУБД), или же традиционных реляционных СУБД (РСУБД). Информация в ООСУБД хранится в виде объектов, что позволяет представлять сложные объекты более непосредственным образом, нежели в РСУБД, и использовать при хранении ту же модель данных, что и в объектно-ориентированных языках программирования. На основе анализа рассмотренных источников можно сделать вывод, что ООСУБД целесообразно применять в специализированных программных системах, обрабатывающих очень большое количество информации (порядка петабайтов), например для узкоспециальных задач в области физики высоких энергий, в то время, как для СЭД более высокий приоритет имеет возможность комплексирования с другими информационными системами организации, при том, что объемы данных не так высоки и, как правило, измеряются тера-, а не петабайтами, вследствие чего для СЭД предпочтительнее использовать ОРСУБД. Для всего комплекса административных информационных систем ЦЕРН в качестве единой СУБД был выбран Oracle, поэтому для СЭД EDH также используется ОРСУБД Oracle (объектные расширения были добавлены в нее, начиная с версии 8), что упрощает комплексирование с остальными административными системами ЦЕРН.

Упрощенная структура хранилища данных ПОЭД СЭД EDH показана на рис. 6.

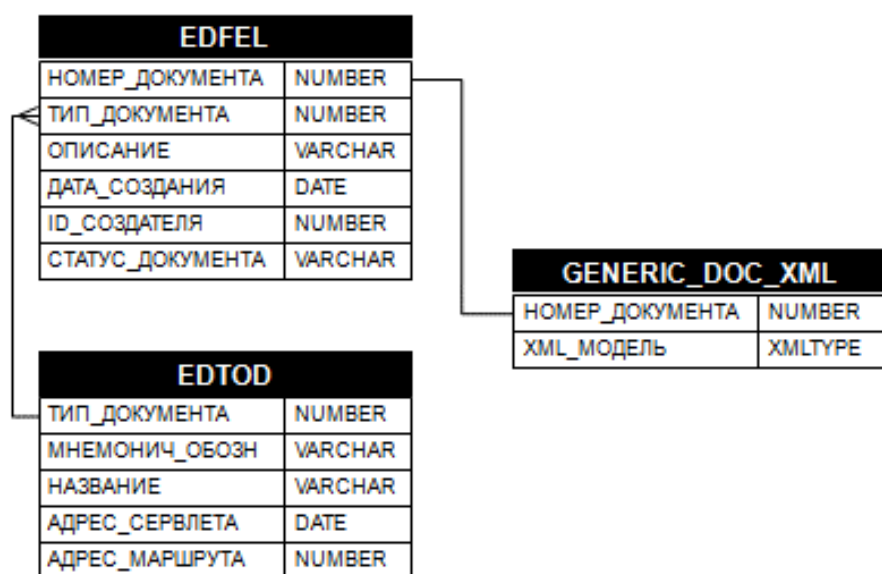


Рис. 6. Упрощенная схема хранилища данных ПОЭД

В таблице EDFEL хранятся общие реквизиты документов, входящие в состав Дублинского ядра. Таблица EDTOD, связанная с EDFEL отношением «один ко многим» содержит информацию обо всех поддерживаемых СЭД бизнес-процессах и соответствующих им типах документов. Наличие данной таблицы позволяет ПОЭД определить, какой сервлет, какие сервисные классы и какую процедуру маршрутизации необходимо использовать для тех или иных документов. Наконец, таблица GENERIC_DOC_XML предназначена для хранения экземпляров документов в виде XML-деревьев. Для сохранения в ОРСУБД реквизитов, имеющих комплексный тип данных, используется рассмотренный в третьей главе подход сохранения частичного набора атрибутов. Для каждого комплексного типа

данных в СЭД EDH имеется объект Java, содержащий атрибуты комплексного типа данных, и программный интерфейс для загрузки объектов данного типа (комплексных объектов) из корпоративных СУБД через интерфейс Enterprise Java Beans. Все эти объекты (в совокупности составляющие так называемую «библиотеку базовых объектов») наследуются от одного интерфейса. Поэтому для сохранения атрибутов комплексных объектов можно использовать механизмы инкапсуляции и полиморфизма ООП, для чего в используемый интерфейс введены два новых метода, отвечающих за сохранение частичного набора атрибутов объекта в XML-модели и загрузку объекта из XML-модели:

```
public void сохранитьXML(String XML-модель); // сохранить объект в XML
public DataCarrier загрузитьXML(String XML-модель); // загрузить объект
```

Системное программное обеспечение EDH включает в себя СУБД Oracle 10g фирмы Oracle и пакет серверных приложений Oracle Application Server 10g той же фирмы. Пакет серверных приложений состоит из веб-сервера на основе ядра известного веб-сервера Apache, контейнера корпоративных приложений языка Java (Oracle Application Server Container for J2EE), а также утилит для балансировки загрузки серверов (load balancing), контроля и администрирования. СУБД Oracle установлена в корпоративном варианте (Enterprise Edition) и включает в себя средство Oracle Text для создания полнотекстовых индексов и поиска по ним.

Одним из важнейших компонентов интерфейса СЭД является поисковая система. В рамках данной работы исследованы поисковые системы Oracle Text и Microsoft FAST, каждая из которых позволяет учитывать структуру документа, заданную XML-деревом модели экземпляра. Для реализации поисковой системы СЭД EDH была выбрана система полнотекстового поиска Oracle Text в сочетании со сделанной в ЦЕРН системой генерации отчетов ART, дающей возможность легко создавать сложные запросы к СУБД путем задания структуры запросов и таблиц данных при помощи XML. Архитектура разработанной поисковой системы показана на рис. 7.

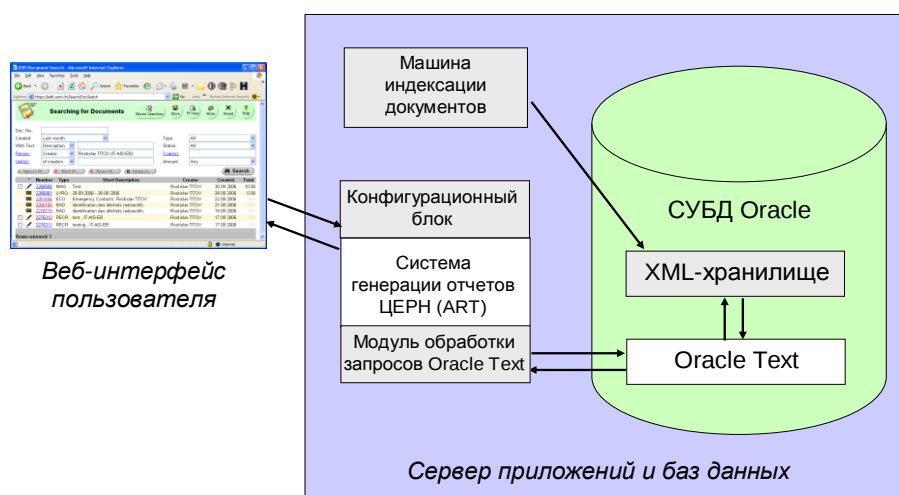


Рис. 7. Архитектура поисковой системы СЭД EDH.

Анализ запросов к поисковой системе в ЦЕРН показал, что в большинстве случаев пользователи интересуются документами, созданными за последние 3-6 месяцев. Для дополнительного повышения быстродействия поисковой системы при поиске недавно созданных документов были созданы отдельные усеченные

поисковые индексы, размеры которых примерно в 15 раз меньше размера полного индекса, благодаря чему полнотекстовые запросы Oracle Text к такому индексу выполняются существенно быстрее.

На рис. 8 показан внешний вид пользовательского интерфейса разработанной поисковой системы СЭД EDH. Форма в верхней части экрана предназначена для ввода элементов описания для поиска документов и формирования запросов (при этом допускаются сложные запросы, в которых группы элементов описания объединяются логическими признаками И, ИЛИ, НЕ). Результирующий список содержит идентификатор документа, тип документа, имя создателя и дату создания, описание документа и (для финансовых документов) итоговую стоимость. Иконографическое изображение в левой части таблицы показывает, на какой стадии утверждения находится документ (полностью утвержден, отклонен, находится в процессе подписания, находится в процессе создания). Поисковая система позволяет осуществлять поиск как по общим реквизитам, так и по уникальным реквизитам документов – при выборе типа документа из соответствующего списка в интерфейсе появляются дополнительные поля для осуществления запросов с использованием уникальных реквизитов.

The screenshot shows the EDH Document Search interface in a Mozilla Firefox browser. The page has a header with the EDH logo and navigation links. Below the header, there are two main sections for search parameters. The first section, 'Параметры запроса', includes fields for document number, creation time, content, employee, and department. The second section, 'Параметры запроса (2)', includes dropdowns for document type, status, and code, and a text field for the total sum. Below these sections is a 'Сложный запрос - критерии' section with a table of criteria. At the bottom, there is a table of search results with columns for document number, type, description, creator, creation date, and sum.

№ док-та	Тип	Краткое описание	Создатель документа	Дата создания	Сумма
4524359	DAI	Imac 27 inch from ART Computer	Heather GRAY	03.12.2010	2,005.00
4524208	MAG	(BAAN4524208) HP desktop computer	David COCKERILL	03.12.2010	684.00
4523693	TNRQ	Removal of old computer screen	Iulia PASCU	03.12.2010	N/A
4522965	MAG	(BAAN4522965) emory upgrade for radmon computer	Pierre Alain GANDER	02.12.2010	63.00
4522411	MAG	(BAAN4522411) Printer 3525 / spare	Christine LEROY-JONCKX	02.12.2010	954.00
4520291	MAG	(BAAN4520291) COMPUTER FOR LINUX	Gustavo SEGURA MILLAN	30.11.2010	706.00
4519391	DAI	Computer fan and power supply items	Olivier RAGINEL	29.11.2010	434.94
4518072	MAG	(BAAN4518072) Dell Optiplex Computer	Nitesh SONI	27.11.2010	1,058.50
4517635	MAG	(BAAN4517635) Standard desktop computer HP 8000 Elite	Francesco Paolo LA TORRE	26.11.2010	684.00
4517548	DAI	Computer Equipment	Dimitar MLADENOV	26.11.2010	188.80
4516819	DAI	Computer Equipment	Dimitar MLADENOV	25.11.2010	76.69
4516564	MAG	(BAAN4516564) Standard desktop computer DELL Optiplex7	Hyunkwan SEO	25.11.2010	1,058.00
4514910	MAG	Toner printer	Claude LAMBOLEY	24.11.2010	781.00
4514400	MIS	Computer Equipment	Christian ISNARD	23.11.2010	N/A

Рис. 8. Внешний вид поисковой системы СЭД EDH. Пример поиска документов, содержащих в кратком описании слово «computer» или содержащих в описании слово «printer» и имеющих стоимость от 500 до 1000 франков. Примечание. Краткие описания найденных документов не переведены на русский язык, поскольку они не имеют отношения к интерфейсу поисковой системы.

Внедрение корпоративных информационных систем (КИС), к которым, безусловно, относятся СЭД, сопряжено со значительными временными и денежными затратами, в связи с чем их срок службы может составлять 5-15 лет и более. Важной характеристикой таких систем является масштабируемость, то есть возможность оперативной коррекции характеристик системы при увеличении числа поль-

зователей и/или объемов входных/выходных потоков. В СЭД EDH масштабируемость обеспечивается путем кластеризации и балансировки загрузки. Входной поток данных обрабатывается аппаратным маршрутизатором, равномерно распределяющим запросы между несколькими серверами приложений, работающих под управлением ОС Linux и составляющих кластер EDH. В свою очередь на каждом из серверов приложений запросы обрабатываются программным распределителем загрузки, входящим в состав сервера приложений Oracle Application Server 10g и обеспечивающим дальнейшее равномерное распределение входного потока данных между несколькими независимыми виртуальными машинами Java, установленными на каждом из серверов, а обращение к данным обеспечивается ОПСУБД Oracle, обладающей собственными средствами кластеризации (рис. 9).

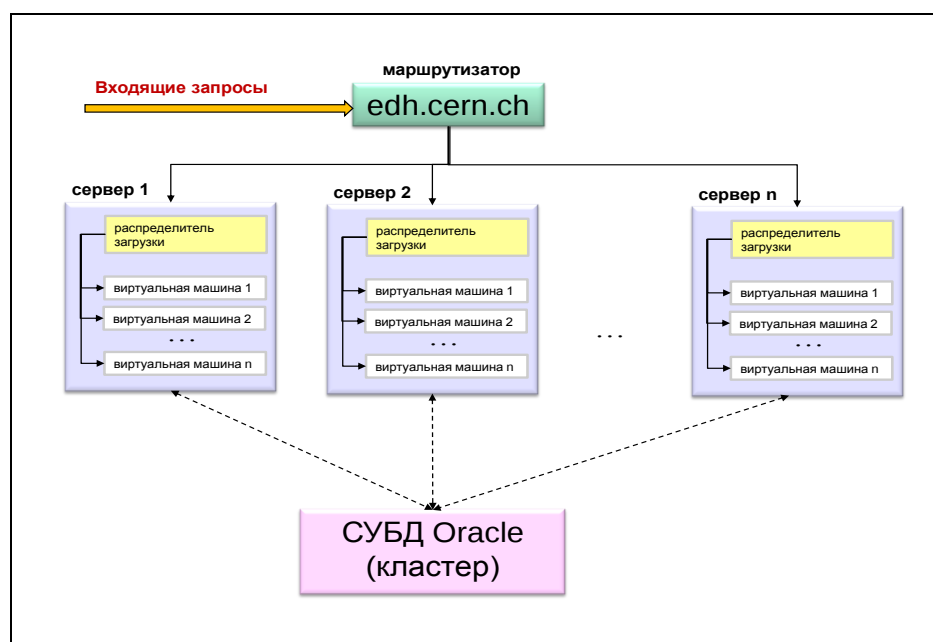


Рис. 9. Обеспечение масштабируемости СЭД EDH

Максимально целесообразное число виртуальных машин Java на каждом из серверов определяется количеством доступных процессоров и оперативной памяти. При увеличении объемов входных/выходных потоков в такой архитектуре достаточно увеличить число виртуальных машин Java, либо увеличить число серверов приложений в кластере. При этом системный сбой на какой-либо виртуальной машине или сервере затрагивает лишь небольшое количество пользователей, а новые пользователи автоматически распределяются между оставшимися серверами и виртуальными машинами до тех пор, пока неисправность не будет устранена, благодаря чему обеспечивается не только масштабируемость, но и бесперебойность работы системы.

На графиках рис. 10 показан ежегодный рост числа пользователей и рост числа электронных документов, ежегодно создаваемых в системе EDH. За этот период аппаратная конфигурация системы была расширена от одного сервера приложений Sun и одной виртуальной машины Java до кластера из 6 серверов Linux, на каждом из которых работает по 4 виртуальных машины Java. Несмотря на существенное увеличение числа пользователей и количества обрабатываемых документов, скорость обработки запросов в системе не ухудшилась, что практически подтверждает правильность выбора такой архитектуры.

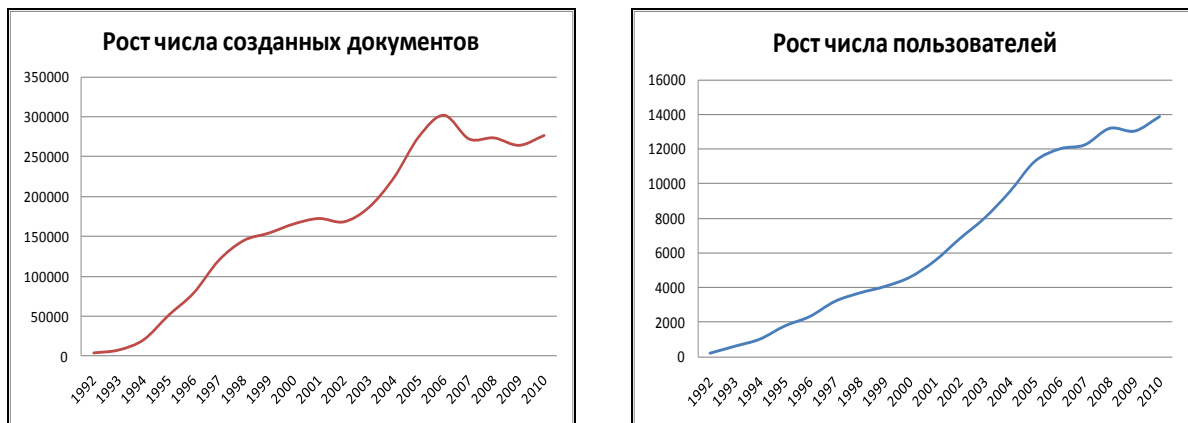


Рис. 10. Ежегодная динамика числа пользователей и числа созданных документов СЭД EDN

Переносимость характеризует возможность переноса программы с одной платформы на другие с минимальными её переделками и является достаточно важной характеристикой КИС, так как, с одной стороны, позволяет легко адаптировать систему к изменяющейся во времени аппаратной базе предприятия, что увеличивает срок службы КИС.

Для приблизительной оценки сокращения времени разработки можно использовать модель оценки стоимости разработки программного обеспечения COSOMO II. По классификации COSOMO II СЭД EDN можно считать «полуразделенным проектом», а факторам стоимости, присутствующим в модели COSOMO, исходя из оценки проекта, можно присвоить нормальный рейтинг. В этом случае формулы для расчета модели COSOMO принимают следующий вид:

$$E = 3,0 \left(\frac{LOC}{1000} \right)^{1,12}, \quad (1)$$

$$D = 2,5(E)^{0,35}$$

где:

E – затраты на разработку в человеко-месяцах;

D – время разработки в хронологических месяцах;

LOC – количество строк исходного кода.

В качестве оценки количества строк исходного кода LOC будем использовать средние значения эквивалентного количества строк исходного кода при автоматизации бизнес-процессов с использованием разработанной подсистемы обработки электронных документов и без нее. Тогда:

$$E_n = 3,0 \left(\frac{1928}{1000} \right)^{1,12} = 6,258, \quad D = 2,5(6,258)^{0,35} = 4,75$$

$$E_c = 3,0 \left(\frac{6319}{1000} \right)^{1,12} = 32,891, \quad D = 2,5(32,891)^{0,35} = 7,65$$

Таким образом, сокращение затрат на автоматизацию новых бизнес-процессов с использованием предлагаемой методологии составляет:

$$F = \frac{D_c}{D_n} = \frac{7,65}{4,75} = 1,61$$

Оценка по методу COSOMO показывает, что применение подсистемы обработки электронных документов позволяет сократить время автоматизации электронных документов на 61%. На практике в ЦЕРН наблюдалось сокращение времени разработки в 2 и более раз при автоматизации бизнес-процессов с помощью нового подхода. При этом разработка электронных документов с использованием подсистемы обработки электронных документов и моделированием документов на языке XML может вестись силами менее квалифицированного персонала (например, студентов), что обеспечивает дополнительное снижение затрат на автоматизацию новых бизнес-процессов.

ЗАКЛЮЧЕНИЕ

В результате проведенных исследований решена актуальная научно-техническая задача разработки системы электронного документооборота и создания программного комплекса для автоматизации управления научно-техническими проектами.

Основные результаты, полученные при выполнении данной работы, заключаются в следующем:

1. Проведен системный анализ научно-методологических и технологических аспектов решения исследуемой задачи, с помощью которого определены и обоснованы принципы построения систем электронного документооборота, обеспечивающих автоматизацию бизнес-процессов и управление потоками работ в научно-технической сфере.

2. Разработаны формальные модели электронных документов, описывающие сложные иерархические структуры данных с использованием XML-деревьев.

3. Разработаны модели экземпляров электронных документов с выделением общих и уникальных реквизитов, применимые для широкого класса документов. Модели инвариантны к типам данных, что обеспечивает универсальность их применение в различных системах электронного документооборота.

4. Определены способы группирования общих реквизитов документов, позволяющие компенсировать несогласованность структур данных, что упрощает хранение и поиск документов. Правильность формирования электронного документа обеспечивается посредством задания XML-схемы для модели экземпляра и используемых типов данных.

5. Разработаны и исследованы алгоритмы обработки и хранения в СУБД моделей документов, содержащих комплексные типы данных, в результате чего выбран способ сохранения частичного набора атрибутов комплексных типов данных.

6. Разработаны методы построения систем семантического полнотекстового поиска электронных документов на базе системы Oracle Text, которые позволяют осуществлять высокоскоростной поиск документов как по общим реквизитам, так и по реквизитам, уникальным для некоторого типа документов. При этом допускается использование сложных запросов, в которых условия поиска объединяются логическими признаками.

7. На основе проведенных исследований модернизирована система электронного документооборота ЦЕРН, использующая декларативное описание бизнес-логики в модели структуры документа, что позволило более чем в два раза со-

кратить время на автоматизацию новых бизнес-процессов. С использованием разработанной архитектуры в ЦЕРН создано свыше 2000000 документов.

8. С помощью разработанной системы электронного документооборота реализован бизнес-процесс маркировки радиоактивных отходов, являющийся исходным этапом процедуры обеспечения радиационной безопасности работы в ЦЕРН.

9. Разработанная методология применения XML для описания иерархических документов использована при создании системы отслеживания событий в системе управления выполненной стоимостью для проекта Большой Адронный Коллайдер. Данная система позволяет отслеживать события, происходящие в системе управления выполненной стоимостью, оценивать их влияние на график стоимости работ и дает информацию по принятию управленческих решений.

Основные публикации по теме диссертации

Список в изданиях, внесенных в перечень ВАК Минобрнауки РФ

1. *Фергюсон Дж. М., Пурвис Дж., Йонге Ю. Д., Маттесон Д., Титов Р.* Средства генерации административной отчетности: современный подход // Приборы и системы. Управление, контроль, диагностика – 2005. № 8, стр. 58-63.
2. *Д. Маттесон, Р. Титов, Дж. Пурвис, Ю. де Йонге, Дж. Фергюсон.* Система работы с электронными документами ЦЕРН // Приборы и системы. Управление, контроль, диагностика – 2005. № 9, стр. 62-67.
3. *Ю. Де Йонге, П. Бонналь, Дж. Фергюсон, Д. Маттесон, Дж. Пурвис, Р. Титов.* Система управления выполненной стоимостью для проекта Большой Адронный Коллайдер // Приборы и системы. Управление, контроль, диагностика – 2005. № 10, стр. 60-65.
4. *Р.Н. Титов.* Применение расширяемого языка разметки XML при создании адаптируемых интерфейсов Web-приложений // Приборы и системы. Управление, контроль, диагностика – 2006. № 2, стр. 62-65. ISSN 2073-0004.
5. *Голиков К.О., Леонова Н.М., Титов Р.Н.* Проектирование предметно-ориентированных языков программирования на примере системы управления документами ЦЕРН // Естественные и технические науки – № 6, 2009 – с. 530-539. ISSN 1684-2626
6. *Голиков К.О., Леонова Н.М., Титов Р.Н.* Подход к решению проблемы разработки и поддержки электронных документов на примере системы административного документооборота ЦЕРН // Системы высокой доступности – 2011, №1, стр. 48-55.

Публикации в других изданиях

7. *Ильиных Г. С., Васильев Н. П., Титов Р. Н.* Система обработки файловых вложений для системы электронного документооборота ЦЕРН (EDH) // Научная сессия МИФИ-2003, сборник научных трудов, т.13, Конференция "Молодежь и наука". Компьютерные науки. Информационные технологии, стр. 116-117.
8. *Букреев Ю. В., Титов Р. Н., Чернышев Ю. А.* Разработка модуля ввода и отображения информации для системы стоимостного анализа проектов (EVM) // На-

учная сессия МИФИ-2003, Сборник научных трудов, т.10, Телекоммуникационные и новые информационные технологии в образовании, стр. 118-119.

9. *R. Titov. Practical Use of XML – CERN Computing Seminar* [электронный ресурс] // CERN – 2004 // URL: <http://indico.cern.ch>

10. *Фергюсон Дж., Йонге Ю. Д., Пурвис Дж., Матиесон Д., Титов Р. Н.* Административные информационные системы ЦЕРН как способ устранения несоответствий между потребностями пользователей и возможностями имеющихся на рынке систем // Научная сессия МИФИ-2005. Сборник научных трудов XII Всероссийской научной конференции «Проблемы информационной безопасности в системе высшей школы». М.: МИФИ, 2005, стр. 86-87.

11. *Гречкин Т. Ю., Матиесон Д., Титов Р.Н., Щукин Б.А.* Практические вопросы разработки электронных документов в системе электронного документооборота EDH // Научная сессия МИФИ-2005, Сборник научных трудов, т.10, Телекоммуникационные и новые информационные технологии в образовании, стр. 22-24.

12. *Мыхтарянц Ю.А., Чернышев Ю.А., Титов Р. Н., Матиесон Д.* Применение технологии Spring в системе электронного документооборота ЦЕРН (EDH) // Научная сессия МИФИ-2005, Сборник научных трудов, т.10, Телекоммуникационные и новые информационные технологии в образовании, стр. 24-26.

13. *Д. Матиесон, Р.Н. Титов.* Использование среды World Wide Web и технологий J2EE при автоматизации корпоративного документооборота // Научная сессия МИФИ-2006, сборник научных трудов, т. 10, Телекоммуникации и новые информационные технологии в образовании, стр. 17-18. ISBN 5-7262-0633-9.

14. *Батуашвили И., Титов Р., Матиесон Д.* Модуль администрирования системы электронного документооборота ЦЕРН // Научная сессия МИФИ-2006, сборник научных трудов, т.12, Информатика и процессы управления. Компьютерные системы и технологии, стр. 14-15.

15. *А. Цыганов, Р. Титов, Д. Матиесон.* Модуль резервного сохранения и восстановления документов системы электронного документооборота ЦЕРН // Научная сессия МИФИ-2006, сборник научных трудов, т.12, Информатика и процессы управления. Компьютерные системы и технологии, стр. 37-38.

16. *Запевалов И.М., Титов Р.Н., Матиесон Д.* Повышение интерактивности пользовательских интерфейсов веб-приложений на примере системы электронного документооборота ЦЕРН // доклад на Московской международной телекоммуникационной конференции молодых ученых и студентов «Молодежь и наука». Секция 4-1. Технологии разработки программных систем – 2008.

17. *Запевалов И.М., Титов Р.Н., Матиесон Д.* Применение передовых поисковых технологий на примере системы электронного документооборота ЦЕРН // доклад на Московской международной телекоммуникационной конференции молодых ученых и студентов «Молодежь и наука». Секция 4-1. Технологии разработки программных систем – 2008.

18. *Голиков К.О., Титов Р.Н.* Использование предметно-ориентированного кода и оценка меры его сопровождаемости // «Математика. Компьютер. Образование». Сборник тезисов XVII международной конференции – Москва-Ижевск, 2010 – с.103. ISBN 978-5-93972-797-6

19. *R. Titov. Workflows and business process modelling for CERN's Electronic Document Handling system* // Materials of the 1st JINR/CERN International School on Grid and Management Information Systems // Joint Institute for Nuclear Research, Dubna, Russia – 2010