snowball.tartarus.org/algorithms/russian/stemmer.html. 13. *Keleberda I., Repka V., Biletskiy Y.* 2006. Building learner's ontologies to assist personalized search of learning objects. ICEC 2006. P. 569-573.

Поступила в редколлегию 16.05.2008

Шамша Борис Владимирович, канд.техн.наук, профессор кафедры ИУС ХНУРЭ. Научные интересы: разработка эффективных методов кластеризации в ИУС. Адрес: Украина, 61166, Харьков, пр. Ленина 14, тел. 702-14-51.

Шатовская Татьяна Борисовна, канд.техн.наук, доцент кафедры ПОЭВМ ХНУРЭ. Научные интересы: Data mining, Web mining, Text mining. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. 702-14-46.

Гуд Анастасия Юрьевна, аспирантка кафедры ИУС ХНУРЭ. Научные интересы: разработка эффективных методов классификации в информационных управляющих системах. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. 702-14-51.

УДК 519.23

В. М. ЛЕВЫКИН, Е. А. МОСПАН

РАЗРАБОТКА МОДЕЛИ ТЕХНОЛОГИИ ФОРМИРОВАНИЯ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ В WEB-ОРИЕНТИРОВАННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Описываются результаты исследования существующих технологий формирования электронных документов в WEB-ориентированных информационных системах, а также их обобщенная модель. В связи с выявленными недостатками подобных технологий разрабатывается модифицированная модель технологии формирования электронных документов с учетом специфики WEB-ориентированных информационных систем. Модифицированная технология предполагает формирование электронных документов на основании шаблона, представленного описательными языками, и разметки, а также выдачи выходного документа в формате, определенном пользователем.

1. Введение

В настоящее время электронные документы занимают важное место в области передачи информации и взаимодействия в человеко-машинных системах. Современные системы управления документами (Document Management System - DMS)[1], системы управления контентом (Content Management Systems - CMS)[2] и информационные системы (ИС) различных классов немыслимы без электронных документов. В первом случае электронные документы составляют основу организации документооборота, во втором являются одним из основных продуктов функционирования системы, который может выступать не только вариантом представления данных системы, но и юридическим документом. В связи с этим задача формирования и поддержки выходных документов возникает практически у каждого разработчика ИС.

2. Актуальность исследования

Формирование электронного документа в ИС связано с конкретной функциональной задачей или бизнес-функцией, а именно с входными и выходными данными, относящимися к ним. Процесс внесения изменений данных, необходимых для документа, неразрывно связан с изменениями требований к самой задаче. Однако подобные изменения не означают внесения корректировок в структуру электронного документа, которая в свою очередь может изменяться вне зависимости от реализации требуемой задачи. Добавление новых требований к структуре чаще всего определяется заказчиком. Исходя из этого, выходной электронный документ необходимо рассматривать как совокупность его структуры и данных, которые определяются состоянием системы. Под структурой документа будем понимать его шаблон, который определяет форматирование и представление выходного документа. Под данными будем понимать набор входных и выходных данных функциональной задачи, которые необходимо отразить в создаваемом документе.

В рамках формирования документов для функциональных задач, в системах DMS и CMS возникает проблема: в каком формате должен быть сформирован выходной документ? Выбирая определенный формат в традиционных системах, эту проблему можно решить посредством формирования жестких требований к системному программному обеспечению (СПО), которое должно быть установлено на машине клиента. В связи с тем, что WEB-ориентированные ИС имеют клиент-серверную архитектуру (в большинстве случаев с тонким клиентом), требования к СПО, установленному на ПК клиента, должны быть минимальны. Отличительной особенностью WEB-ориентированных ИС является также и то, что доступ к ним осуществляется с различных платформ, которые определяются не только операционной системой, но и типом устройства, через который осуществляется доступ (стационарные рабочие станции, карманные компьютеры, мобильные телефоны и прочее). Следует отметить, что количество операционных систем для этих устройств, а соответственно и прикладного ПО, очень разнообразно. Для ПК это операционные системы Windows, Unix и Linux и другие. Для мобильных – это Symbion, Windows Mobile, IMacOS и другие. Предположим, что некоторая WEB-ориентированная ИС формирует электронные документы для своих клиентов только в языке разметки Rich Text Format (RTF). Это означает, что на каждом ПК клиента должно быть установлено соответствующее программное обеспечение, которое позволит просмотреть и при необходимости отредактировать сформированный электронный документ. Однако подобное СПО может отсутствовать для ОС определенного класса или просто может быть не установлено на машине клиента. Таким образом, клиенту не предоставляется возможность выполнять свои функции, что противоречит основным принципам построения WEB-ориентированных ИС, один из которых говорит о доступности подобной системы из любого места. В связи с этим задача формирования документов на основании шаблона усложняется тем, что выходной документ должен генерироваться в требуемом пользователю формате.

3. Существующие технологии формирования электронных документов

На основании сказанного выше можно сделать вывод о том, что клиенту необходимо предоставить возможность выбора формата электронного документа, который формирует та или иная ИС. Рассмотрим способы, при помощи которых решаются подобные задачи на современном этапе развития информационных технологий.

Одним из способов является хранение документов на сервере в различных форматах. В рамках такого подхода возможно наличие шаблона документа, представленного требуемым форматом, преобразование документа осуществляется посредством использования объектной модели этого формата. При этом на сервере должно быть установлено соответствующее программное обеспечение, которое обеспечивает доступ к объектной модели документа (например, COM-технология, которая обеспечивает доступ к документам, создаваемым в пакете Microsoft Office). Недостаток такого подхода заключается в том, что подобное программное обеспечение будет обязательным для сервера, что ведет к проблемам, касающимся совместимости, поддержки и удорожания системы.

Другой способ – использование широкого инструментария для генерации документов в различных форматах, т. е. система располагает различными средствами генерации отчетов в том или ином формате. Наличие шаблона определяется спецификой выбираемого программного обеспечения для реализации подобной задачи. При этом для генерации отчета в некотором формате должен существовать программный модуль, обеспечивающий эту функциональность. В этом случае количество подобных модулей равно количеству форматов, которые необходимо поддерживать. Это приводит к сложностям как в процессе разработки, так и в период поддержки системы. Среди свободно распространяемых средств формирования выходных документов следует выделить такие: IText, JasperReports, JFreeReport, iReport и другие [3].

Также возможно применение XSLT-преобразований. Шаблон представляется в формате регламентируемым FOP. Динамические данные вносятся в него посредством расширения стандартной библиотеки дополнительных программных модулей, что ведет к похожим недостаткам, указанным в предыдущем подходе. Преимуществом использования подобного подхода к генерации отчетов является то, что посредством XSLT-преобразований можно получать документы в различных форматах [4].

Существующие технологии по формированию документов T представим в виде следующей модели:

$$T = < (P, D, S, O) \Phi_S^P, \Phi_S^D, \Phi_O^S >, \quad (1)$$

где P – категория, представляющая шаблоны документов; D – категория, представляющая входные и выходные данные системы; S – категория, описывающая модель структуры шаблона документа; O – категория, представляющая выходные документы системы; Φ_S^P - функтор, отображающий категорию шаблонов документов в модель его структуры; Φ_S^D - функтор, отображающий категорию данных системы в элементы модели структуры документа; Φ_O^S - функтор, отображающий категорию модели структуры электронного документа в выходной документ системы.

Недостатками рассмотренных технологий формирования документов являются:

- ограниченный набор форматов, которыми можно представлять шаблоны документов;
- ограниченный набор форматов, которыми могут быть представлены выходные документы;
- в некоторых случаях необходима установка дополнительного программного обеспечения на сервер;
- зависимость от ОС для определенных технологий формирования выходных документов;
- использование специфических форматов представления документов-шаблонов;
- необходимость применения другой технологии формирования документов при возникновении задачи формирования документа в неподдерживаемом формате.

4. Модифицированная технология формирования электронных документов

Для исключения этих недостатков необходима модифицированная технология формирования электронных документов. Рассмотрим основные предпосылки к разработке подобной технологии формирования документов. Прежде всего, необходимо обеспечить гибкость создания документов-шаблонов. Для решения подобной задачи целесообразно использовать современные языки разметки описательного типа, которые поддерживаются мощными текстовыми процессорами. Такими языками разметки являются, например, RTF, HTML, OpenDocument, OpenXML и другие[5]. Создаваемая технология должна обеспечить возможность разработчику использования произвольного языка разметки описательного типа, что позволит гибко подходить к вопросу создания документа-шаблона. При этом использование множества языков разметки описательного типа предполагает разработку унифицированной модели, которая описывает структуру документа. Для такой унифицированной модели структуры документа должны быть разработаны методы, которые позволяют гибко изменять структуру документа и вносить в неё данные, соответствующие актуальному состоянию ИС.

В качестве формата выходных документов эффективно использовать языки разметки описательного и процедурного типов. Если для описательных языков разметки в разрабатываемой технологии формирования документов предусмотрена унифицированная модель структуры документа, то для языков процедурного типа необходимо разработать модель с абсолютным позиционированием элементов. Это связано со спецификой представления данных в языках разметки процедурного типа, к которым относятся PDF, TeX, LaTeX и другие[6]. Следует отметить, что модель структуры с абсолютным позиционированием элементов должна автоматически создаваться на основании унифицированной модели структуры документов.

Учитывая указанные выше предпосылки и недостатки существующих технологий представления и преобразования структур электронных документов, мы предлагаем модифицированную технологию формирования документов на основании шаблонов, которая позволяет значительно сократить временные и трудовые затраты на этапе создания и внесения изменений выходных документов ИС. Общая структура такой модифицированной технологии представлена на рисунке.

Преимуществами такой технологии являются:

- уменьшение времени форматирования электронного документа;
- возможность использования различных форматов для подготовки документа-шаблона;
- простота внедрения и поддержки нового формата электронного документа;
- возможность предопределения пользователем формата выходного документа;
- отсутствие необходимости установки дополнительного ПО на сервер и на ПК клиента.

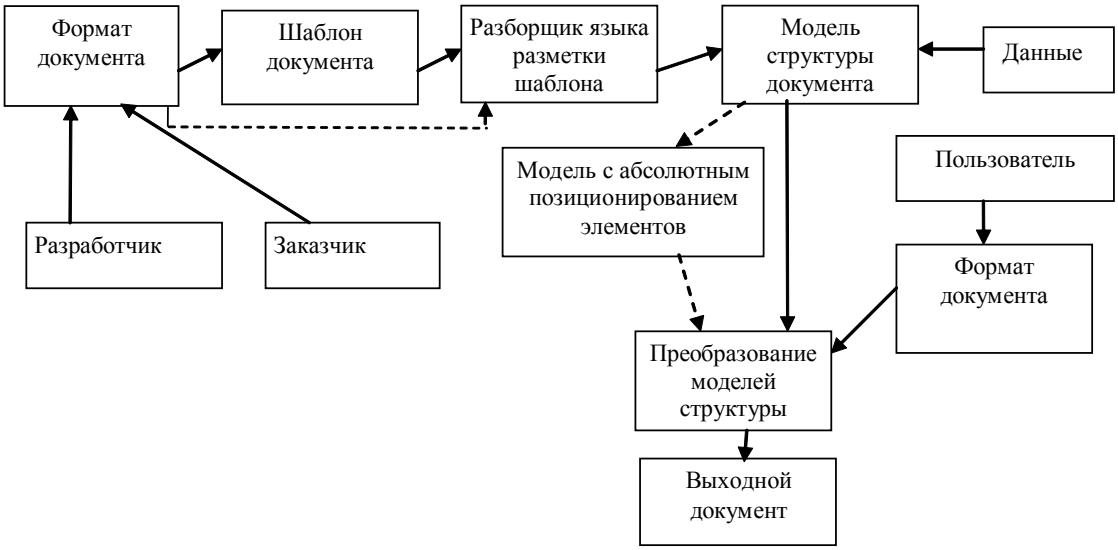
Тогда модель модифицированной технологии формирования документов T' можно представить следующим образом:

$$T' = < (I,\{S_n\},D,M_{ld},M_{pd},O_d,O_p)\{\Phi^{I}_{S_n}\},\{\Phi^{S_n}_{M_{ld}}\},\Phi^{D}_{M_{ld}},\Phi^{M_{ld}}_{M_{pd}},\Phi^{M_{ld}}_{O_d},\Phi^{M_{pd}}_{O_p} >, \quad (2)$$

где I – категория, представляющая языки разметки описательного типа, которые могут быть использованы для представления шаблона документа системы; $\{S_n\}$ - множество категорий, представляющих модели структуры документов языков разметки, которые входят в категорию I, при этом n – это количество языков описательного типа, которые входят в I; D – категория, представляющая входные и выходные данные системы; M_{ld} – категория, представляющая унифицированную модель структуры документа, используя которую можно представить любую модель, входящую в $\{S_n\}$; M_{pd} – категория, представляющая модель структуры документа с абсолютным позиционированием элементов, относительно листа фиксированного размера; O_d - категория, представляющая языки разметки описательного типа, которые могут быть использованы в качестве формата выходного документа; O_p - категория, представляющая языки разметки процедурного типа, которые могут быть использованы в качестве формата выходного документа; $\{\Phi^{I}_{S_n}\}$ - множество функторов, отображающее категорию языка разметки описательного типа в соответствующую ему категорию модели структуры документа; $\{\Phi^{S_n}_{M_{ld}}\}$ - множество функторов, отображающее категорию модели структуры документа, представленного описательным языком разметки, в категорию унифицированной модели структуры документа; $\Phi^{D}_{M_{ld}}$ - функтор, отображающий категорию данных системы в элементы унифицированной модели структуры документа; $\Phi^{M_{ld}}_{M_{pd}}$ - функтор, отображающий категорию унифицированной модели структуры документа в категорию модели структуры с абсолютным позиционированием элементов; $\Phi^{M_{ld}}_{O_d}$ - функтор, отображающий унифицированную модель структуры документа в категорию языков описательного типа; $\Phi^{M_{pd}}_{O_p}$ - функтор, отображающий модель структуры с абсолютным позиционированием элементов документа в категорию языков процедурного типа.

Формирование выходного документа посредством модифицированной технологии осуществляется в виде следующих этапов:

- разработчик выбирает формат документа-шаблона из множества поддерживаемых языков разметки описательного типа, представленного категорией I;
- осуществляется переход к структуре документа-шаблона, представленного выбранным языком разметки описательного типа, благодаря одному из функторов множества $\{\Phi^{I}_{S_n}\}$;
- осуществляется переход к унифицированной модели структуры документы, которая представлена категорией M_{ld} на основании одной из категории множества $\{S_n\}$ и функтора из множества $\{\Phi^{S_n}_{M_{ld}}\}$;
- в структуру создаваемого выходного документа вносятся изменения, соответствующие актуальному состоянию ИС с использованием категории D, которая представляет входные и выходные данные функциональной задачи, а также функтор $\Phi^{D}_{M_{ld}}$;
- создается выходной документ с использованием функторов $\Phi^{M_{ld}}_{O_d}$ или $\Phi^{M_{pd}}_{O_p}$ на основании выбранного пользователем формата выходного документа из категорий O_d или O_p, которые представляют поддерживаемые языки разметки описательного и процедурного типов соответственно;



Общая структура модифицированной технологии формирования электронных документов

- создается модель структуры документа с абсолютным позиционированием элементов, которая применяется для получения выходного документа посредством функтора $\Phi_{O_p}^{M_{pd}}$, в случае генерации выходного документа в формате языка разметки процедурного документа с использованием функтора $\Phi_{M_{pd}}^{M_{ld}}$.

5. Выводы

Таким образом, разработана модифицированная технология формирования электронных документов в WEB-ориентированных ИС. Эта технология обеспечивает возможность использования в качестве документа-шаблона произвольного языка разметки описательного типа. В качестве формата предполагается применение одного из языков разметки описательного или процедурного типа. Кроме того, была расширена кортежная модель технологии создания выходных документов. Реализация предложенной модели может быть осуществлена с использованием методологии объектно-ориентированного проектирования.

Список литературы: 1. *Глинских А.* Мировой рынок систем электронного документооборота. http://www.iteam.uublications/it/section_64/article_2582. 2. *Content* management system - http://en.wikipedia.org/wiki/Content_management_system. 3. *Open* Source Charting & Reporting Tools in Java - http://www.java-source.net/open-source/charting-and-reporting. 4. *Clark, J* (ed). XSL Transformations (XSLT). Version 1.0. W3C Recommendation 16 November 1999- http://www.w3.org/TR/xslt. 5. *Goldfarb, C. F.* The SGML Handbook. Oxford University Press, Oxford, UK. 1990. 6. *Andrew T.* Young. Principles of LaTeX formatting. 2005-2006 - http://mintaka.sdsu.edu/GF/bibliog/latex/principia.html.

Поступила в редколлегию 30.03.2008

Левыкин Виктор Макарович, д-р.техн.наук, профессор, зав. кафедрой ИУС ХНУРЭ. Научные интересы: разработка корпоративных ИС, синтез сложных ИС. Увлечения: автотуризм, видеофильмы. Адрес: Украина,61022,Харьков, ул. Чичибабина, д.2,кв.83, тел. 705-40-91.

Моспан Евгений Александрович, аспирант кафедры ИУС ХНУРЭ, начальник отдела Java ООО «ПрофИТсофт». Научные интересы: технологии проектирования WEB-ориентированных информационных систем. Увлечения и хобби: футбол, снукер. Адрес: Украина, 61142, Харьков, ул. Гарибальди, 7, кв. 63, тел. 7178-109 (дом), +38097-457-84-01 (моб).