Уфа : УГАТУ, 2009 Вестник УГАТУ • Управление, ВТ и И Т. 13, № 2 (35). С. 188–197

УДК 004.658

В. В. МИРОНОВ, Г. Р. ШАКИРОВА

ТЕХНОЛОГИЯ ПЕРСОНАЛИЗАЦИИ ДОКУМЕНТОВ ФОРМАТА OPEN OFFICE XML НА ОСНОВЕ XSL-ТРАНСФОРМАЦИИ

Обсуждаются вопросы персонализации электронных документов путем XSL-трансформации XML-данных. Объясняется, как использовать текстовый процессор Microsoft Word для создания сложных XSL-спецификаций трансформации и классы платформы .NET для автоматизации программирования и использования спецификаций. *Электронные документы*; *персонализация*; *XML-технологии*; *XSL-трансформация*; *.NET*

В последнее время мировое IT-сообщество проявляет все больший интерес к политике ведущих корпораций, нацеленной на перевод программного обеспечения на платформу XML. Пример тому – широко разрекламированная полемика по вопросу стандартизации нового формата Open Office XML корпорации Microsoft.

Начиная с 2002 г., компания Microsoft планомерно вводит XML в линейку своих продуктов – сначала XML-ориентированным стал Word (в 2002 г.), а затем и остальные компоненты пакета Office (с 2003 г.). Формат XML-Office постоянно перерабатывался, и каждая новая версия приносила в него нечто новое. В конечном итоге Microsoft в 2007 г. полностью изменила Office – XML стал его основой, а не просто «приятным» дополнением. Этот формат стал известен под названием Open Office XML. Он включает в себя не только документы Word, но и электронные таблицы Excel, рисунки Visio, презентации PowerPoint и автономные формы InfoPath.

Такое нововведение способствовало тому, что кардинально изменился подход к программированию документов Office. Отпала необходимость (хотя возможность и осталась) сохранения исходных документов с расширением xml – документ изначально представлен в XML-формате. Это привело к появлению новых программных и инструментальных механизмов их создания и модификации, а также появлению новых и коренной переработке известных задач, связанных с такими документами.

Одной из актуальных задач программирования документов Office является их персонализация. Персонализация предполагает подстановку известных персональных данных в некоторую заготовку и формирование таким образом конечного документа, «заточенного» под конкретного пользователя [1].

Задача персонализации документов Word уже обсуждалась авторами в предыдущих публикациях [1]. Вместе с тем, предложенные там решения работоспособны только применительно к Word 2003. Новая структура документов Word позволяет существенно расширить возможности этих решений и распространить их возможности (с некоторой доработкой) на более сложные задачи персонализации. Рассмотрим в данной статье, как можно решить задачу персонализации и адаптировать уже предложенные решения к формату Word 2007.

1. СТРУКТУРА ДОКУМЕНТА WORD 2007

Документ Word 2007 представляет собой сжатый zip-архив, называемый пакетом (package), внутри которого размещены отдельные файловые компоненты (parts) в формате XML [2]. Рядовой пользователь зачастую и не догадывается о такой сложной организации документа, с которым он работает. На первый взгляд такой документ практически ничем не отличается от своих предшественников. Исключение составляет только расширение – docx, свидетельствующее о возможной «причастности» XML (x – XML).

В отличие от традиционного формата doc, docx можно назвать только условно закрытым. С одной стороны, невозможно расценивать документ такого формата просто как XML-документ. Его код, открытый в текстовом редакторе, зашифрован. Значит, казалось бы, не-

Контактная информация: (347) 272-89-81

возможно внести изменения в такой документ, минуя процессор Word.

С другой стороны, можно пойти другим путем. В первоначальном варианте документ формата docx представляет собой «неявный» архив. Его невозможно распаковать стандартными архиваторами, такими, как, к примеру, WinRar. Чтобы преобразовать такой неявный архив в явный, нужно к исходному полному имени файла добавить расширение zip. Тогда с помощью любого доступного архиватора можно просмотреть его структуру (рис. 1). Поскольку каждый компонент пакета – XML-документ с открытым кодом, его можно просмотреть в любом браузере и модифицировать в любом редакторе.

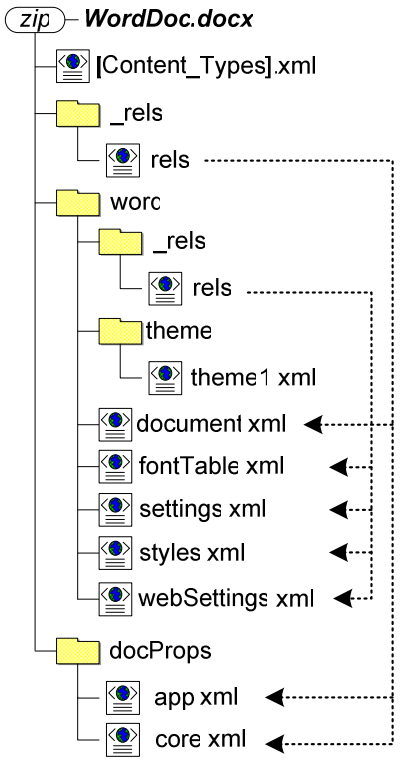


Рис. 1. Структура сжатого zip-пакета Word 2007

Структура zip-пакета представлена как совокупность самостоятельных компонентов в виде XML-документов. Эти компоненты можно разделить на две группы – реляционные и типизованные. Каждый из них описывает определенные составляющие конечного документа.

В пакете представлено два вида реляционных компонентов. Первый – файл [Content_Types].xml – содержит ссылки на основные разделы документа. Второй – раздел _rels – задает взаимосвязи отдельных разделов документа друг с другом или элементов внутри одного раздела.

Основное содержимое документа задается в типизованном разделе word. В нем представлены файлы, содержащие описание свойств документа, доступных шрифтов, стилей и др. Здесь же располагается и главный компонент документа – файл document.xml, содержащий информационный контент конечного документа. К нему так или иначе привязаны все остальные компоненты пакета-документа Word.

Другое типизированное содержимое – встроенные в документ объекты. К ним относятся, к примеру, графика и мультимедиа. В пакете Word такие объекты хранятся в своем исходном формате.

Структура файла document.xml повторяет структуру сохраненного в формате XML документа Word 2003 за одним лишь исключением. Корневым элементом стал w:document, а все пространства имен «обновились» до более поздних версий. Кроме того, разметка была дополнена несколькими системными элементами и атрибутами, раскрывающими свойства документа и его внутреннего кода.

Рассмотрим, как решаются задачи персонализации на уровне обновленного формата.

2. ЗАДАЧА ПЕРСОНАЛИЗАЦИИ ДОКУМЕНТОВ WORD 2007

Целью персонализации является формирование электронных документов, содержащих персональные данные определенного пользователя. Такой процесс предполагает наличие шаблона конечного документа, в котором предусмотрены правила подстановки пользовательских данных.

Процесс персонализации является в большей степени серверным. Это значит, что формирование документов предпочтительнее выполнять в обход процессора Word, чтобы максимально снизить нагрузку на сервер. Выполнение персонализации на локальном уровне очевидно целесообразно в тех случаях, когда нужно сформировать достаточно большое количество документов с одними и теми же персональными данными. В остальных случаях персонализованные документы можно построить и в среде Word.

В работе [1] предложена серверная технология персонализации электронных документов на основе XSL-трансформации. К пользовательским данным, представленным в XML-формате (XML-базе реквизитов), применяются XSL-спецификации преобразования. Такие спецификации представляют собой макеты конечных документов и содержат WordML-разметку формируемого документа и XSL-

инструкции подстановки пользовательских реквизитов.

Процесс персонализации можно разделить на два этапа: программирование XSL-спецификаций персонализации и собственно генерацию персонализованных конечных документов.

Каждый из этапов по-своему специфичен применительно к формату Open Office XML. Это связано, прежде всего, с тем, что документ такого формата разбит на несколько составных частей, а потому важно еще на этапе проектирования определиться с тем, какие именно части должны быть персонализованы. Это, с одной стороны, облегчает процесс персонализации, поскольку трудоемкость и масштаб программирования сужаются, а с другой – усложняет его, поскольку уже отработанные технические решения, применявшиеся к формату Word 2003, здесь не работают. Последнее утверждение справедливо уже потому, что ранее использованные подходы предполагали, что документ Word неделим, его метаданные (шрифты, стили, параметры и др.) и собственно информационное содержимое не отделены друг от друга. Это противоречит главному принципу версии 2007 – отделению данных от их представления.

Рассмотрим подробнее каждый из этапов персонализации.

3. ПРОГРАММИРОВАНИЕ XSL-СПЕЦИФИКАЦИЙ ПЕРСОНАЛИЗАЦИИ

Обобщенная схема программирования XSL-спецификаций персонализации документов Word 2007 приведена на рис. 2.

Особенность программирования спецификаций персонализации документов Word 2007 определяется их компонентной архитектурой. Теоретически, можно персонализовать каждый XML-компонент сжатого пакета. Важно отметить, что XML-структура доступна изначально и нет необходимости сохранять документ в XML-формате (хотя и это возможно). Это позволяет относительно просто строить XSL-спецификации для персонализации каждого компонента.

Однако на практике задачи персонализации решаются скорее на уровне информационного наполнения документа, чем его оформления и параметров. В структуре пакета Word такое содержимое задается в файле document.xml. Поэтому целесообразно «запускать» спецификации персонализации только для этого компонента пакета. Иными словами, при работе с документами Word 2007 имеет место частичная персонализация. Такой подход вынуждает несколько пересмотреть технологию проектирования инструкций персонализации.

Использование XSL-инъекций для создания XSL-спецификаций персонализации

Отличительной особенностью документов Word, независимо от версии, является возможность их представления в формате XML. Для этого среда Word располагает встроенной схемой WordML (Wordprocessing Markup Language), в соответствии с которой размечается исходный документ. В результате закрытый формат документа становится полностью открытым как для просмотра, так и для модификации разработчиком или сторонними программами, отличными от Word.

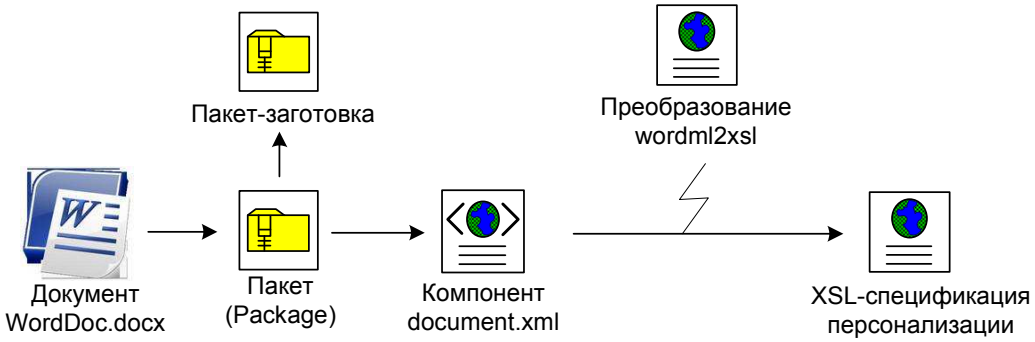


Рис. 2. Схема программирования XSL-спецификаций персонализации документов Word 2007

Такая особенность позволяет существенно упростить процедуру персонализации документа. Достаточно просто определить WordML-разметку конечного документа и задать в ней правила подстановки пользовательских реквизитов.

WordML-разметку конечного документа можно построить непосредственно в среде Word. Для этого разработчик создает новый документ, определяет его структуру и правила форматирования, а потом сохраняет в формате XML. В результате формируется WordML-разметка, которая должна быть сгенерирована в персонализованном документе.

В работе [1] предложено представлять пользовательские данные в формате XML в виде XML-базы реквизитов. Процесс персонализации сводится к преобразованию этих данных в WordML-разметку конечного документа, т. е. одних XML-данных в другие. Такое преобразование может быть выполнено с помощью XSL-спецификаций трансформации. Тогда инструкции подстановки пользовательских реквизитов – это XSL-инструкции подстановки XML-данных с помощью соответствующих выражений XPath-адресации.

Чтобы упростить программирование XSL-спецификаций персонализации, XSL-инструкции подстановки пользовательских данных можно вводить прямо в среде Word. Такие программные «вкрапления» называются XSL-инъекциями.

Так, к примеру, можно построить XSL-спецификацию персонализации для документа, приведенного на рис. 3. Документ представляет информацию о поставках заданного поставщика: данные о самом поставщике, его поставках и товаров в этой поставке. Для придания документу формы XSL-спецификации, он дополнен XSL-инъекциями подстановки пользовательских данных value-of и XSL-инструкцией объявления входного параметра для определения данных конкретного поставщика.

В ходе исследования обработки XSL-инъекций в документах Word 2003 была разработана XSL-таблица стилей, задающая правила преобразования WordML2XSL. Она позволяет создать XSL-спецификацию трансформации, копируя заданную WordML-разметку и превращая текстовые XSL-инъекции в программные XSL-инструкции.

Применительно к Word 2007 преобразование WordML2XSL должно быть выполнено на уровне документа document.xml. Поскольку его структура практически полностью повторяет WordML-разметку документа Word, то к нему применимо предложенное преобразование WordML2XSL. Вместе с тем, XSL-спецификация WordML2XSL требует дополнительной корректировки применительно к формату Word 2007. Для этого нужно объявить в таблице стилей пространства имен WordML 2007 и удалить все «следы» Word 2003.

Дополнительная сложность преобразования WordML2XSL применительно к Word 2007 связана с расстановкой пробелов. В Word 2003 корректная расстановка пробелов определялась атрибутом «xml:space» корневого элемента. В результате «пробельные» правила распространялись на весь документ целиком.

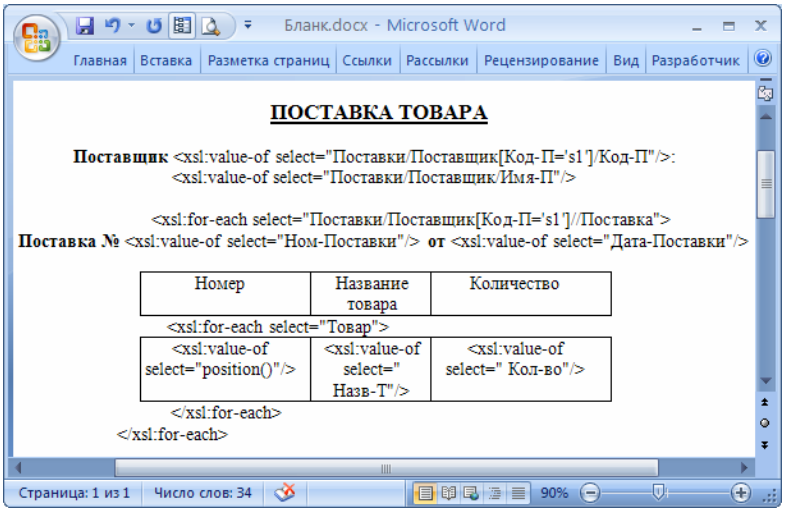


Рис. 3. Пример документа Word 2007 с XSL-инъекциями

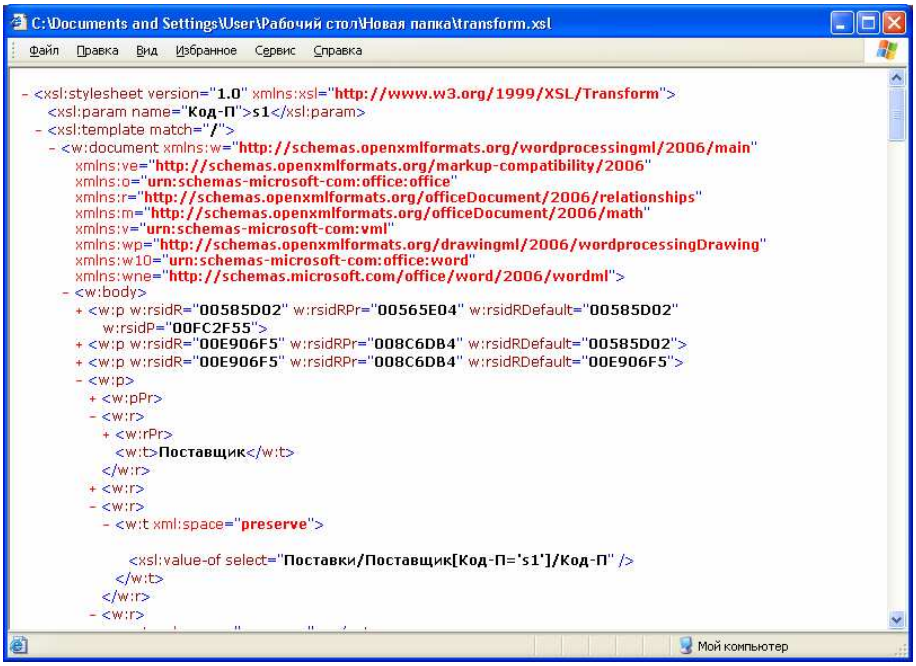


Рис. 4. Пример результата преобразования WordML2XSL

В WordML 2007 правила расстановки пробелов задаются для каждого отдельного текстового фрагмента, поэтому «склеивание» XSL-инъекции, разбитой на несколько текстовых фрагментов, чревато утратой каких-то нужных или появлением лишних пробелов. В этой связи функция построения XSL-инструкции из фрагментированной XSL-инъекции была дополнена, с одной стороны, механизмом удаления лишних пробелов внутри самой XSL-инъекции, а с другой – добавлением пробелов до и после сформированной XSL-инструкции.

В результате, применяя преобразование WordML2XSL к файлу document.xml для документа, приведенного, к примеру, на рис. 3, можно получить корректную XSL-спецификацию персонализации. Отображение полученной спецификации в окне браузера Internet Explorer приведено на рис. 4.

Корректность спецификации персонализации нужно протестировать на конкретных данных. Так, к примеру, можно применить преобразование к данным, приведенным на рис. 5. Поскольку в данном примере спецификация преобразования параметризована, то ее использование нужно сопроводить передаваемым значением, к примеру, «s1». В результате вместо XSL-инструкций value-of должны быть подставлены конкретные значения (s1, Смит и т. д.), произойдет персонализация компонента document.xml.

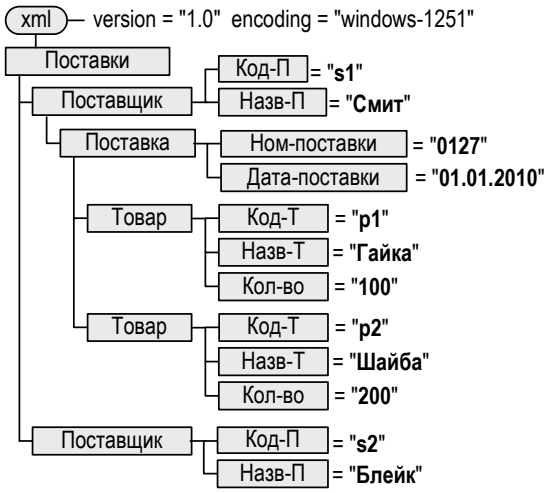


Рис. 5. Пример модели данных

Создание пакета-заготовки

Вместе с тем персонализация одного компонента – еще не персонализация всего документа. Персонализованный компонент необходимо добавить в сжатый пакет Word. Для этого помимо XSL-инструкции персонализации на сервере должен храниться уже готовый пакет Word. В этот пакет подставляется персонализованный компонент, который потом передается пользователю.

Чтобы персонализованный компонент не конфликтовал с уже имеющимся в пакете, целесообразно в размещенном на сервере пакете заведомо удалить переменную часть – файл

document.xml. Такой пакет называется пакетом-заготовкой.

При подстановке персонализованного файла document.xml в пакет-заготовку формируется конечный персонализованный документ формата Word 2007. Так, если компонент document.xml, приведенный на рис. 4, добавить в пакет-заготовку, то будет получен документ со сведениями о поставках поставщика с кодом s1 (рис. 6).

Автоматизация программирования XSL-стилей персонализации

Чтобы упростить программирование XSL-стилей персонализации, его можно автоматизировать. Для этого используется серверный сценарий, выполняющий формирование пакета-заготовки и XSL-спецификации персонализации. В этой связи возникает проблема выбора программной технологии, с помощью которой можно построить сценарий.

С появлением нового формата компания Microsoft предложила и программную технологию его обработки – Open Office XML SDK [3]. В рамках этой технологии доступна библиотека классов .NET, позволяющих выполнять любые манипуляции с документами Word 2007. Эти классы дополняют доступные классы платформы .NET Framework по работе с архивными пакетами, каковым, по сути, и является документ Word 2007.

Основным классом, который используется при работе с документами, является класс Package. Каждый экземпляр этого класса представляет контейнер, в котором может храниться несколько объектов данных. Основным физическим форматом этого класса является zip-файл.

При построении пакета-заготовки разработчик загружает в экземпляр класса Package документ Word, находит и удаляет из него часть document.xml. Модель программного кода, выполняющего эти операции, приведена на рис. 7, *а*.

Важно отметить, что для выполнения сценария нет необходимости преобразовывать документ в явный архив. Сценарий по умолчанию будем рассматривать документ как сжатый пакет, несмотря на отличное от привычных архивных файлов расширение docx.

Еще одна задача, которая может быть выполнена при помощи класса Package, связана с формированием XSL-спецификации персонализации. Задача усложняется тем, что разработчик вводит XSL-инъекции в документ, а не в отдельный его компонент. Преобразование WordML2XSL выполняется именно на уровне такого компонента. Значит, нужно извлечь компонент из архива и применить к нему заданное преобразование.

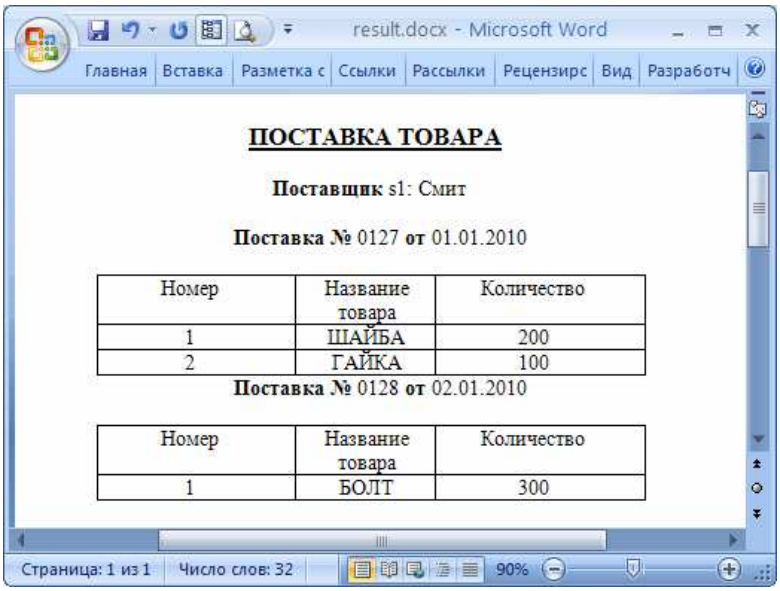
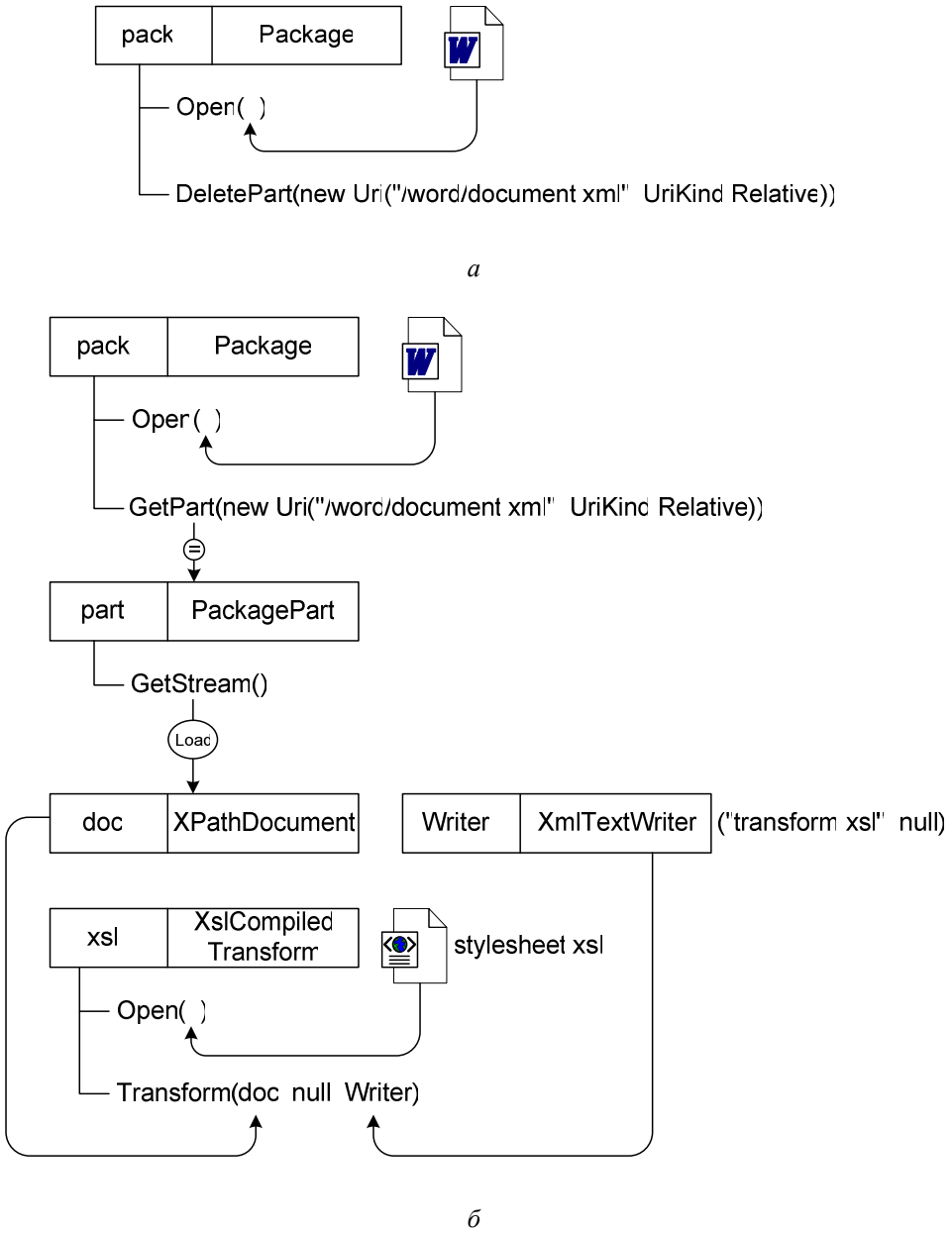


Рис. 6. Пример персонализованного документа Word 2007



а

б

Рис. 7. Пример модели сценария программирования XSL-инструкций персонализации

На уровне сценария документ эта задача решается следующим образом (рис. 7, *б*). Документ с XSL-инъекциями загружается в экземпляр класса Package. Здесь к нему применяется метод GetPart, который извлекает из пакета часть document.xml. Полученные XML-данные помещается в экземпляр DOM-объекта, и к нему применяется преобразование WordML2XSL. Результатом преобразования и будет искомая спецификация персонализации.

3. СЦЕНАРИИ ГЕНЕРАЦИИ ПЕРСОНАЛИЗОВАННЫХ ДОКУМЕНТОВ

Наличие на сервере пакета-заготовки и XSL-спецификации персонализации позволяет сформировать и передать пользователю конечный документ, заполненный его персональными данными. Обобщенная схема этого процесса приведена на рис. 8.

Сценарий персонализации на сервере выполняется по запросу пользователя. Он на программном уровне применяет XSL-спецификации персонализации к пользовательским XML-данным. В результате трансформации формируется персонализованный компонент document.xml, который добавляется в размещенный на сервере пакет-заготовку. Полученный в итоге пакет docx отправляется пользователю.

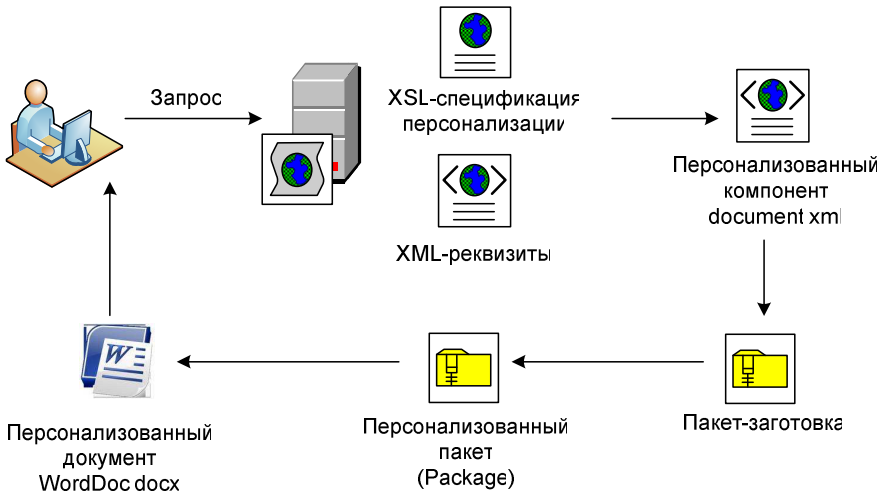


Рис. 8. Обобщенная схема формирования персонализованного документа Word 2007

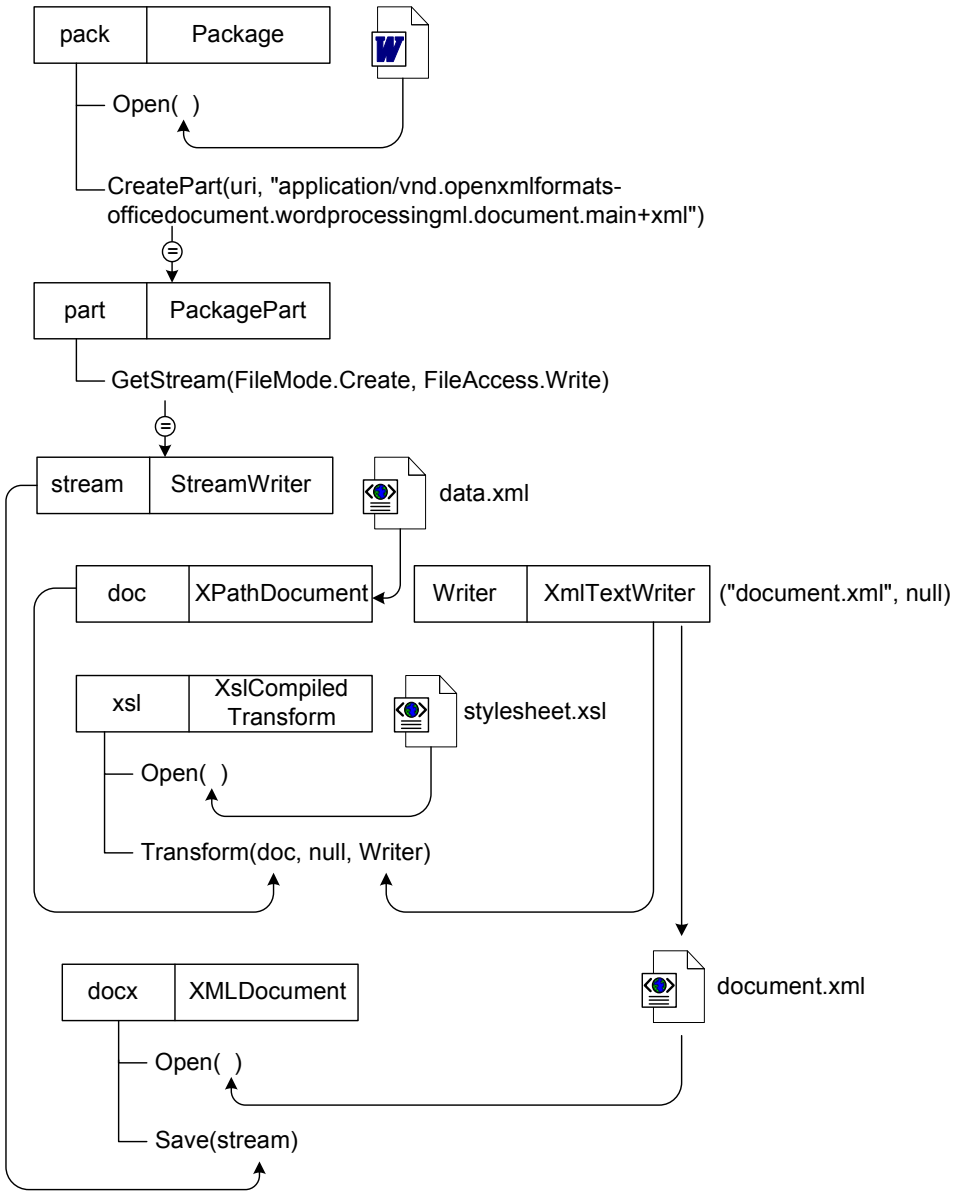


Рис. 9. Пример модели сценария персонализации документов Word 2007

Поскольку основные операции персонализации выполняются на уровне пакета Word и его составляющих, сценарий должен быть реализован на базе .NET классов Package. Соответствующая программная модель приведена на рис. 9.

Вместе с тем, на начальном этапе персонализации сценарий оперирует стандартными DOM-объектами – загружает в них пользовательские данные и XSL-таблицу стилей. Далее на их основе выполняется XSL-трансформация и формируется персонализованный компонент.

На следующем этапе в экземпляр класса Package загружается размещенный на сервере пакет-заготовка. Сформированный персонализованный компонент с помощью метода CreatePart добавляется в пакет-заготовку и передается пользователю.

4. ОБЕСПЕЧЕНИЕ БЕЗОПАСНОСТИ ПЕРСОНАЛИЗАЦИИ ДОКУМЕНТОВ WORD 2007

Тот факт, что документ Word 2007 представляет собой zip-архив, применительно к безопасности является одновременно его преимуществом и недостатком.

При открытии документа каждый компонент zip-архива проходит тщательную проверку на валидность. Если какой-то файл архива поврежден (corrupted), то при попытке открытия документа в среде Word будет предложено его восстановить. Кроме того, сама zip-упаковка документа предусматривает контроль по суммам CRC, что исключает возможность искажения данных на этапе передачи файлов пользователю. Таким образом, можно сказать, что новый формат Open Office XML обеспечивает многоуровневую поддержку безопасности своего содержимого.

С другой стороны, присутствует и ряд уязвимостей, связанных с все тем же форматом Open Office XML. Во-первых, пакет-заготовка хранится на сервере. Ничего не стоит подменить один из его компонентов другим, в лучшем случае просто поврежденным. Во-вторых, поскольку компоненты хранятся в открытом XML-формате, можно их разрушить и вручную, к примеру, просто удалив из разметки один из тегов, что приведет к тому, что документ будет некорректным. Наконец, можно просто удалить один из компонентов из пакета-заготовки, и документ так же будет поврежден. Для борьбы с этим можно, к примеру, на программном уровне выполнять проверку наличия всех необходимых компонентов в передаваемом пользователю пакете и контролировать их корректность и валидность по крайней мере по нескольким параметрам.

Еще одна уязвимость связана с XSL-спецификацией персонализации. Спецификация также хранится в открытом формате, который легко можно изменить. В итоге трансформация может привести к результату, совершенно отличному от ожидаемого, и, наконец, повреждению конечного пакета-документа. Чтобы избежать этого, можно выполнять проверку применяемой таблицы стилей еще на уровне ее загрузки в экземпляр DOM-объекта.

ЗАКЛЮЧЕНИЕ

1. Формат Open Office XML – новый формат представления офисных документов, предложенный компанией Microsoft. Его XML-направленность позволяет выполнять любые операции с документами Word, Excel, PowerPoint и др. Так, документы формата Word могут быть обработаны с помощью серверных сценариев, не требующих запуска процессора Word. Одной из актуальных задач, которые можно решить по такой технологии, является персонализация документов.

2. Документ Word 2007 представляет собой zip-архив (пакет), содержащий XML-файлы – компоненты пакета. Информационное содержание документа заключено только в одном из компонентов – файле document.xml. Следовательно, персонализация такого документа предполагает создание архива-заготовки и XSL-спецификации персонализации компонента document.xml.

3. Программирование XSL-спецификации персонализации на начальном этапе может быть выполнено в среде Word. Для этого разработчик создает документ и определяет его структуру. Правила расстановки пользовательских данных задаются в том же документе с помощью введенных в тексте XSL-инструкций – XSL-инъекций. Формирование пакета-заготовки и извлечение компонента document.xsl с XSL-инъекциями может быть выполнено на уровне .NET класса Package.

4. Формирование персонализованного документа предполагает персонализацию компонента document.xml и его последующую подстановку в пакет-заготовку. Полученный в результате персонализованный документ Word 2007 передается пользователю. Для реализации этой функциональности так же используется класс Package, который записывает в пакет-заготовку преобразованный файл document.xml.

5. Безопасность персонализации документов Word 2007 может быть выполнена, с одной

стороны, с помощью механизмов защиты и восстановления, предусмотренных самим форматом Open Office XML, а с другой, требует определенных действий со стороны разработчика приложения персонализации.

СПИСОК ЛИТЕРАТУРЫ

1. **Миронов В. В., Шакирова Г. Р., Яфаев В. Э.** Информационная технология персонализации электронных документов Microsoft Office в Web-среде на основе XML // Вестник УГАТУ: научн. журн. Уфимск. гос. авиац. техн. ун-та (сер. «Управление, вычислительная техника и информатика»). 2008. Т. 10, № 2(27). С. 112–122.

2. **Новиков И.** Office 2007. Новая платформа разработки // PC Magazine [Электронный ресурс] (http://pcmag.ru/solutions/sub_detail.php?ID=6348&SUB_PAGE=0).

3. Open XML Developer [Электронный ресурс] (http://openxmldeveloper. org).

ОБ АВТОРАХ

Миронов Валерий Викторович, проф. каф. автоматизир. систем упр-я. Дипл. радиофизик (Воронежск. гос. ун-т, 1975). Д-р техн. наук по упр. в техн. сист. (УГАТУ, 1995). Иссл. в обл. иерархич. моделей и ситуац. управления.

Шакирова Гульнара Равилевна, ст. преп. той же каф. Дипл. инженер по АСОИУ (УГАТУ, 2005). Канд. техн. наук по матем. и прогр. обеспечению вычисл. машин, комплексов и комп. сетей (УГАТУ, 2008). Иссл. в обл. XML-технологий.