УДК 004.057.3:004.912

Сопоставление структур форматов документов odt и docx

И.Н. Магомедов

Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова

Современные текстовые редакторы и процессоры предоставляют пользователю множество удобных и полезных функций и возможностей.

Сейчас мы можем добавлять рисунки, таблицы, формулы, форматировать текст, проверять орфографические ошибки в реальном времени, создавать схемы в текстовом процессоре и множество других возможностей. Это всё очень помогает в написании документации, статей, книг и многого другого. Однако, развитие форматов текстовых документов привело к тому, что из-за большого их разнообразия возникает проблема выбора, какой формат использовать, и как один текст определенного формата конвертировать в другой формат текстового документа.

Проблема связана с тем, что не все форматы текстовых документов имеют открытую документацию, не все форматы поддерживаются в популярных текстовых процессорах и редакторах и не все текстовые процессоры, и редакторы отображают их корректно. Самое важное, что текст написанный в одном формате сложно скопировать в другой формат. Возможности просто скопировать созданный текст одного формата в документ другого формата не существует. Да, есть возможность использовать конверторы текста, но и они не всегда позволяют корректно конвертировать текст.

Сложности появляются ещё и из-за того, что возможности в разных текстовых процессорах отличаются в реализации, и найти какие-то сходства – не тривиальная задача. Имеются различные стандарты, определяющие как реализовывать возможности текстовых процессов. Сопоставим два популярных стандартизованных формата текстовых документов Office Open XML и OpenDocument, а именно форматы .docx и .odt.

Оба этих формата представляют собой zip-архивы. Главное их различие – в структуре самого архива (документа). В docx имеются три папки и один файл, в свою очередь в odt тоже 3 папки, но уже 5 файлов. Папки разделяют документ на отдельные составляющие: папки настроек, метаинформации и список связей документа. И в odt, и в docx имеется файл с перечисление MIME типов содержимого документа. В odt содержится ещё четыре дополнительных файла: файлы контекста, метаинформации, настроек и стилей.

В данной статье рассмотрим внутренние файлы document.xml в docx и context.xml в odt.

В обоих документах первая строка является указанием на версию, на используемую кодировку. Дальше следует указание на начало документа с помощью <w:document> </w:document> в docx и <office:document-content> </office:document-content> в odt. В состав docx разметки document входит один элемент <w:body> </w:body>, в состав odt разметки document входит <office:font-face-decls> с указанием на используемые шрифты, <office:automatic-styles>, в котором располагаются стили для используемых элементов и <office:body>. В docx стили элементов и само содержание располагаются в body и чередуются друг за другом, в odt стили и содержание разделены. Элементы разметки body в docx входят параграфы (<w:p>), run (фрагмент) текста (w:r) и описание страницы (<w:sectPr>), а в odt <office:text> хранящая в себе содержание элементов, частично состоящая из параграфов (<text:p>). В docx вся разметка, входящая в состав body, состоит из параграфов и остальных элементов документа, все они состоят из разметки вида <w: (специальное слово или символ)>. В odt все элементы документа хранятся в office:text, схожесть в разметке имеется не у всех элементов, например, тег <text: (специальное слово или символ)> обозначается текст, нумерованный список и рисунок.

Структура разметки обоих файлов схожа. Можно предположить, что в преобразовании одного формата в другой проблем возникнуть не должно. Но главное отличие заключается в том, что разметка файла docx не хранит полную информацию о логической структуре и визуальном форматировании содержащегося текста, что и вызывает сложности при конвертации текст в другой формат.

Потребность в сопоставление этих форматов документа возникла в связи с разработки программного средства для формирования текстовых документов на основе входных шаблонов. Рассмотрение структуры документов позволит создать программный продукт, позволяющий обрабатывать различные форматы документов и создавать на основе входных шаблонов новые документы различных форматов, что позволит сократить время и расходы при создание документов.