**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 2](#_Toc166077356)

[1 Описание существующих форматов хранения электронных документов 4](#_Toc166077357)

[1.1 EPub 4](#_Toc166077358)

[1.2 FB2 / FB3 6](#_Toc166077359)

[1.3 DjVu 9](#_Toc166077360)

[1.4 PDF 11](#_Toc166077361)

[1.5 DOC / DOCX 15](#_Toc166077362)

[СПИСОК ИСПОЛЬЗОВАНОЙ ЛИТЕРАТУРЫ 16](#_Toc166077363)

ВВЕДЕНИЕ

Информация – неотъемлемая часть нашей жизни. Мы потребляем и обрабатываем ее из различных источников в различных видах и различными способами. Один из видов информации – текст. Он может быть получен из документов, книг, статей, журналов, интернет-сайтов. В цифровую эпоху важным элементом работы с текстом является не только его хранение и передача, но и представление.

В разных структурах и организациях есть свои нормативы для составления документов – стандарты. Они затрагивают любой элемент документа: от заголовков и содержания до таблиц и рисунков. Стандарты позволяют создавать документы в единообразном стиле, избегать коллизий при оформлении того или иного участка (если такое предусмотрено стандартом), а также позволяют описывать структуру документа.

Однако, следовать всем требованиям, накладываемыми нормативами, для людей, создающих документы впервые – задача трудоемкая. Можно запомнить основные правила составления заголовков, абзацев, рисунков, нумерации – этого будет достаточно для создания простого документа. Но могут возникнуть сложности при составлении таблиц, списков (особенно вложенных), математических формул, колонтитулов, списка источников. Также не простым будет процесс редактирования документа для соответствия другому стандарту. В связи с этим становится актуальным

Целью выпускной квалификационной работы является разработка парсера, автоматизирующего процесс представления различных текстовых данных, которые могут быть статьей, отчетом, документом, литературным произведением, в требуемый вид, в зависимости от объекта, где этот текст будет размещен. Поставленная цель требует решения проблемы изменения стиля содержимого, вследствие непреднамеренного упущения какого-либо правила при составлении документа. Для достижения поставленной цели целесообразно использовать язык разметки XML, необходимый для хранения текстовых данных, и файл стилей по аналогии с файлом CSS для HTML.

Новизна:...

1 Описание существующих форматов хранения электронных документов

В контексте цифровых данных целесообразно называть любой цифровой источник текстовой информации, будь то статья, документ, электронная книга, одним словом – электронный документ. Такое обобщенное название условно, ведь для определённых типов документов целесообразней использовать предназначенные для них форматы: для электронных книг – ePub, fb2, djvu, для статей предпочтительней использовать pdf, для документооборота – docx, а для простого текста, не содержащего в себе мультимедиа контента, достаточно txt. Это обусловлено способом взаимодействия с этими данными: создание, редактирование или просто чтение; а также устройством, через которое с ними взаимодействуют создатели или конечные потребители.

Рассмотрим наиболее популярные на сегодняшний день форматы хранения и отображения электронных документов, проанализируем преимущества и недостатки этих форматов.

Большинство из используемых форматов являются контейнерами – представляют собой архив, хранящий основной текст, как правило, в виде XML файла, таблицу стилей, медиаданные и т. п.

1.1 EPub

Electronic Publication (ePub) – открытый формат электронных версий книг. Разработчик – Международный форум цифровых публикаций (IDPF). Вышел в свет в 2007 году и развивается до сих пор – последняя версия 3.3 вышла 25 мая 2023 года.

Данный формат является архивом, содержащим в себе сайт, т. к. накладывает требования на средства его чтения: они должны поддерживать HTML5, JavaScript, CSS, SVG, что, по сути, является требованиям к web-браузерам [1].

Пример содержимого XML файла для формата ePub:

<?xml version="1.0" encoding="UTF-8"?>

<!DOCTYPE html>

<html xmlns="http://www.w3.org/1999/xhtml" xmlns:epub="http://www.idpf.org/2007/ops"

xmlns:ev="http://www.w3.org/2001/xml-events"

epub:prefix="media: http://idpf.org/epub/vocab/media/#">

<head>

<meta charset="utf-8" />

<link rel="stylesheet" type="text/css" href="../css/shared-culture.css" />

</head>

<body>

<section class="base">

<h1>the entire transcript</h1>

<audio id="bgsound" epub:type="media:soundtrack media:background"

src="../audio/asharedculture\_soundtrack.mp3" autoplay="" loop="">

<div class="errmsg">

<p>Your Reading System does not support (this) audio</p>

</div>

</audio>

<p>What does it mean to be human if we don't have a shared culture? What

does a shared culture mean if we can't share it? It's only in the last

100, or 150 years or so, that we started tightly restricting how that

culture gets used.</p>

<img class="left" src="../images/326261902\_3fa36f548d.jpg"

alt="child against a wall" />

</section>

</body>

</html>

К особенностям данного формата можно отнести следующее:

* гибкость отображения. Динамическая разметка позволяет комфортно просматривать документ на любом устройстве: от смартфонов до компьютеров, в зависимости от размеров дисплея;
* конвертируемость. Формат легко преобразовать для чтения в интернете, встроив в сайт или просто открыть через браузер;
* использование закладок;
* выделение отрывков текста и примечания;
* поиск по содержимому. Поиск можно осуществлять из содержащей книгу библиотеки;
* поддержка DRM (Digital rights management англ. – технические средства защиты авторских прав).

Главный недостаток формата ePub связан с информационной безопасностью. Системам чтения, реализующим данный стандарт полностью присуще те же уязвимости, что и web-браузерам. В статье Emma Woollacott от 1 июня 2021 года на сайте portswigger.net говорится: «Используя полуавтоматический испытательный стенд, доступный на GitHub , исследователи обнаружили, что 16 из 97 исследованных систем позволяли ePub сливать информацию о файловой системе пользователя и в восьми случаях извлекать содержимое файла. Злоумышленники, предупреждают они, могут добиться полной компрометации системы пользователя, используя определенные аспекты реализации систем чтения». [2]

Решением проблемы уязвимости является согласие или запрет пользователей на выполнение кода JavaScript в средствах чтения, а также ужесточения требований безопасности.

Формат ePub широко распространён как в российском сегменте Интернета, так и в иностранном. Книги данного формата доступны в Google Play Books, Apple Books, а также в Amazon Kindle.

1.2 FB2 / FB3

FictionBook (англ. – художественная книга) или FeedBook – это формат хранения электронных книг от команды российских разработчиков, во главе с Д. Грибовым, являющегося тех. директором «ЛитРес». FB2 был разработан в 2004 году, последняя версия 2.21 вышла 16 января 2008 года.

По своей структуре это всего лишь XML файл, содержащий только текстовую информацию. Однако с использованием кодировки Base64 можно хранить и изображения.

Фрагмент книги М. Булгакова «Собачье сердце» в формате FB2:

<?xml version="1.0" encoding="UTF-8"?>

<FictionBook xmlns="http://www.gribuser.ru/xml/fictionbook/2.0" xmlns:l="http://www.w3.org/1999/xlink">

<description>

<title-info>

<genre>prose\_classic</genre>

<author>

<first-name>Михаил</first-name>

<middle-name>Афанасьевич</middle-name>

<last-name>Булгаков</last-name>

<id>2bd2e67f-2a82-102a-9ae1-2dfe723fe7c7</id>

</author>

<book-title>Собачье сердце</book-title>

<date>1925</date>

<coverpage>

<image l:href="#cover.jpg"/>

</coverpage>

<lang>ru</lang>

</title-info>

<publish-info>

<book-name>Собачье сердце</book-name>

<publisher>Детская литература</publisher>

<year>2007</year>

<isbn>978-5-08-004244-7</isbn>

<sequence number="0" name="Школьная библиотека"/>

</publish-info>

</description>

<body>

<title>

<p>Михаил Булгаков</p>

<p>Собачье сердце</p>

</title>

<section>

<title>

<p>Глава 1</p>

</title><p>У-у-у-у-у-гу-гуг-гуу! О, гляньте на меня, я погибаю. Вьюга в подворотне ревёт мне отходную, и я вою с ней. Пропал я, пропал. Негодяй в грязном колпаке – повар столовой нормального питания служащих центрального совета народного хозяйства – плеснул кипятком и обварил мне левый бок.</p>

Формат FB2 обладает такими достоинствами как:

* модифицируемость. Достаточно открыть файл в простом текстовом редакторе, однако для его модификации все же нужно знать xml разметку;
* возможность автоматической обработки за счет использования тегов;
* конвертируемость. Формат свободно преобразуется в html либо в pdf, имеет обратную совместимость с его приемником FB3.

Недостатки формата FB2:

* отсутствует поддержка нумерованных и маркированных списков,
* нет сложной верстки текста,
* нет поддержки векторной графики,
* невозможно реализовать защиту авторских прав.

Эти недостатки не существенны, поскольку редко когда художественная литература действительно нуждается в перечисленных элементах. Эти недочеты были исправлены в следующей версии формата FB3, который уже представляет собой zip архив, хранящий текстовый и медиа контент и метаданные по аналогии с ePub. Требования к формату zip-файла и соглашения о его организации прописаны в стандарте ECMA-376 [3]. В отличие от ePub в нем нет поддержки JavaScript, потому FB3 лишен недостатков, связанных с запуском вредоносных скриптов на устройствах пользователей.

Этот формат является открытым и имеет собственный репозиторий на GitHub [4]. Так же на этом ресурсе в открытом доступе есть читатель [5] и редактор [6].

На данный момент можно говорить о большом распространении формата FB2, но только в русскоязычном сегменте Интернета. За рубежом про данный формат практически не известно. Мало популярен, даже в Рунете, и его приемник FB3.

1.3 DjVu

Формат DjVu (фр. déjà vu – «уже виденное») так же предназначен для электронных книг. Начало разработки было положено в 1996 году AT&T Labs, Inc. – научно-исследовательским подразделением американской компании AT&T, специализирующейся на телекоммуникациях. Его разработка была мотивированна распространением отсканированных документов в сети.

Временем начала активного развития интернета можно назвать конец 90-х прошлого века. На тот момент активно поднимался вопрос цифровизации существующей на тот момент литературы. Это было реализовано технически: сканы можно было сохранить в виде картинки в формате jpeg, png. Но эти сканы дорого обходились: «…многие из важных документов имели цветные изображения и фотографии. Чтобы сохранить читаемость текста и качество картинок, приходилось делать сканы в высоком разрешении» [7], их вес мог достигать нескольких мегабайт за страницу. А в то время хранить, и тем более распространять такие тяжелые документы было затратно. Решить данную задачу распознаванием символов, и их в виде текста на тот момент было невозможно – точность была не высокая. Тогда решением могло быть разработка алгоритмов сжатия отсканированных снимков.

К 2001 году закончено создание технологии DjVu. Это позволило сжимать цветные изображения до 40-70 КБ, при оригинальном размере в 25 МБ, текстовые блоки до 15-40 КБ, а сканы древних документов до 100 КБ [8].

Документы DjVu основаны на формате файла контейнера Interchange, который представляет набор чанков – это структура определенного вида. Этот контейнер позволяет хранить данные любого типа: звук, текст, графика. На его основе формат DjVu хранит в цифровом виде информацию в виде изображений.

Сжатие происходит следующим образом: исходное изображение делится на три – передний план, задний план (фон), маска. Маска имеет то же разрешение, что и изображение, но является черно-белой и кодируется определенным образом с использованием алгоритма JB2. Он идентифицирует похожие фигуры на снимке и сохраняет только информацию о расположении данных фигур на нем. Т. о. чем больше в снимке повторяющихся элементов, тем сильнее снимок сожмется. Передний план и фон хранятся в уменьшенном, в сравнении с оригиналом, разрешении и кодируются алгоритмом IW44, при котором изображение теряет в качестве, как и при использовании JPEG, однако, по словам разработчиков, алгоритм IW44 эффективнее его до двух раз.

Преимущества формата DjVu:

* меньший размер документов, за счет лучшего сжатия, чем у основного конкурента PDF;
* возможность производить поиск по тексту, за счет поддержки текстового слоя;
* открытый исходный код.

Однако его преимущество является и основным недостатком, которое затормозило его распространение и, впоследствии, уступило в конкуренции формату PDF. За счет использования алгоритма сжатия JB2 документы имели существенные искажения – похожие символы при низком разрешении сканирования могли интерпретироваться одинаково и заменены при сохранении. Описаны случаи таких замен символов принтерами Xerox Workcentre [9]. Правительственные организации европейских стран запретили использование алгоритма JBIG2 (JB2 является его разновидностью), в целях сохранения отсканированных документов.

В 2016 году проект Internet Archive, курирующий многочисленные проекты по оцифровке книг, заявил о прекращении использования формата DjVu. Основная причина – снижение популярности формата и сложностью поддержки просмотра на Java [10].

К малой популярности можно отнести еще некоторые недостатки:

* ограниченная возможность редактирования,
* ограниченный функциональные возможности,
* отсутствие механизма защиты авторских прав,
* отсутствие защиты от редактирования.

Так как формат хранит лишь отсканированные снимки, то его редактирование может быть осложнено и требовать дополнительных инструментов, особенно при отсутствии текстового слоя. В отличие от главного конкурента PDF, DjVu может хранить только статичную информацию: текст и изображения. Формат не реализует DRM: его нельзя защитить паролем, файл не может быть зашифрован, нет встроенной поддержки цифровой подписи, отсутствуют ограничения на печать, копирование, изменение файлов.

Несмотря на недостатки формат DjVu все еще пользуется популярностью. Однако его развитие остается под вопросом: последняя версия – 26 вышла в апреле 2005 года.

1.4 PDF

Portable Document Format (PDF) – формат хранения документов, разработанный компанией Adobe в начале 90-х. Разрабатывался как средство обмена документами с сохранением форматирования, для просмотра на любом устройстве, не зависимо от операционной системы, а так же печати, в не зависимости от устройства и программного обеспечения.

В первую очередь этот документ решал проблему, когда у пользователей не было необходимых шрифтов для отображения символов, из которых был составлен документ. Их отсутствие не влияло на просмотр документа.

С начала разработки его внедрение в электронный документооборот было медленным. Это было связано с несколькими факторами: первые версии не поддерживали гиперссылки, большой, на момент развития средств передачи документов, размер файлов, но основная причина – проприетарный режим работы с файлами. На тот момент единственным средством просмотра был Adobe Reader, разработанный той же Adobe. Стоимость лицензии составляла 50 долларов, по этой причине только компании, и некоторые частые лица могли позволить себе работать в этой программе.

В 1994 году компания выпустила версию 2.0, которая распространялась бесплатно для любого пользователя. Это главным образом повлияло на популярность формата в Интернете. В 2008 году Adobe опубликовало публичную патентную лицензию по стандарту ISO 32000-1, предполагающей безвозмездное использование PDF. Однако некоторые его спецификации курируются только Adobe, и реализуются только в ее программных продуктах.

Файл PDF является бинарным, хранящий информацию о содержащихся в нем объектах, которыми могут быть текст, изображения, векторная графика, аудио или видео и даже 3D-объекты и их расположением внутри документа. Так же документ позволяет заменять или встраивать шрифты и распространять их вместе с документом, без необходимости установки на устройство пользователя.

Отличительной особенностью PDF от DjVu является тот факт, что PDF возлагает процесс рендеринга документа на программу - читателя, а в DjVu этот процесс выполняет программа - создатель. Это связанно с тем, что PDF кодирует графику и текст в векторизованные данные, и в таком случае программам - читателям необходимо выполнять работу над векторной графикой. В случае же с DjVu он кодирует эти данные как растровые изображения.

PDF-документы имеют две структуры: нелинейную (не «оптимизированную») и линейную («оптимизированную»). Нелинейная структура приводит к меньшему размеру файла, но замедляет доступ к данным, так как элементы, необходимые для создания страниц, распределены по всему файлу. Оптимизированные PDF-файлы, также известные как веб-оптимизированные, разработаны для просмотра в плагинах браузеров без необходимости полной загрузки файла, потому что основные элементы для отображения первой страницы упорядочены в начале файла. По этой причине может возникнуть ситуация, когда меньший по весу PDF файл загружается гораздо дольше файла, большего по объёму.

Перечислим достоинства документа PDF:

* возможность просмотра на большом числе устройств,
* поддержка данных в различных формах (текст, графика, медиа файлы и т.п.),
* реализация различных средств защиты,
* поддержка JavaScript скриптов (опционально),
* стандартизация,
* конвертируемость.

Файлы PDF могут быть просмотрены на любом устройстве конечного пользователя, однако при просмотре на мобильных устройствах, актуальной остается проблема масштабируемости. Ранее мы сказали, что документы данного формата могут предоставлять информацию разного вида. Однако не каждая программа имеет возможность их отобразить – вместо этого файл может просто не открыться, либо вместо проблемного объекта будет отображен другой, если это предусмотрел создатель. Скрипты языка JavaScript поддерживаются отдельной спецификацией PDF, которая не является свободно распространяемой и полноценную поддержку может осуществить ограниченный набор программ.

Встроенные средства защиты документа позволяют, например, защитить документ паролем или сделать доступным только для чтения. В нем так же поддерживается цифровая подпись. Другим немаловажным аспектом в цифровую эпоху является поддержка механизма защиты авторских прав. Приобретенную книгу или иной документ можно просмотреть в определенной программе, введя дополнительную информацию, подтверждающую факт покупки.

Файл преобразуется в различные другие форматы, если это позволяют его настройки. Однако не вся информация может правильно преобразоваться. Проблемы могут возникнуть с текстом. В случае, если на устройстве пользователя отсутствует шрифт, который применялся в PDF, документ может потерять текстовую информацию. То же касается и кодировки. Для успешной конвертации необходимо знать ту кодировку, которая использовалась при создании PDF.

Существует еще одна проблема при работе с документами, связанная с текстом. Он может быть отображен как растровый объект, проще говоря – изображением. Такая ситуация возникает, когда реальный объект сканируется в PDF, без распознавания текста с помощью оптического распознавания символов, а также при намеренном преобразовании в такой вид. Это усложняет работу с данными которые необходимо извлечь из документа.

Несмотря на многочисленные преимущества, PDF также имеет ряд существенных недостатков, помимо тех, которые описаны ранее. Часть из них может негативно сказаться на пользовательском опыте и функциональности документов. Сложности возникают при редактировании PDF, особенно с работой с текстом. Некоторые веб-браузеры позволяют вносить изменения, но только дополняя информацию в файле, например выделив существующие элементы, но не удаляя их.

Еще одна не очевидная проблема связана с информационной безопасностью. Простейшим примером того, как может использоваться документ PDF является фишинг. Злоумышленникам достаточно разместить ссылку на вредоносный ресурс в документе. Но это проблема не связана напрямую с форматом – ссылки могут быть расположены в любом текстовом документе.

PDF-файлы могут быть заражены вирусами, троянами и другими вредоносными программами. В некоторых случаях даже просмотр документа может быть фатальным для жертвы хакеров. Злоумышленники пользуются уязвимостями не только формата, но и программ-читателей. В случае если в читателе включена функция запуска JavaScript скриптов, пользователь может стать жертвой хакеров. Специалист по кибербезопасности Стив Гибсон рекомендует отключать эту функцию, не смотря на пользу, которую она могла бы дать при добросовестном использовании скриптов в документе [11]. Так же описаны случаи, когда сразу при открытии файла вызывалось контекстное меню программы-читателя с предложением сохранить файл PDF. В системе на самом деле сохранялся не PDF a EXE файл, являющийся вирусом.

Большинство известных уязвимостей активно устраняется Adobe, выпуская обновления для софта и обновляя стандарты безопасности. Но те устройства и программы, которые не обновлены до последних версий все еще находятся в группе риска.

Несмотря на описанные недостатки PDF остается основным форматом, который используется в разных видах деятельности человека: от создания статей до электронного документооборота.

1.5 docx

Формат .docx был впервые представлен вместе с выпуском Microsoft Office 2007. Он стал частью стандарта Office Open XML (OOXML), который был разработан компанией Microsoft как открытый формат для хранения документов. Разработка стандарта Office Open XML была начата в 2003 году и была принята как международный стандарт ISO/IEC 29500 в 2008 году.

Технически документ этого формата – это zip архив, содержащий 2 типа файлов: xml и медиафайлы. Первый предназначается может быть представлен двумя форматами .xml и .rels. Они содержат структурированную информацию о форматировании, стилях, текстового содержимого. Медиафайлы чаще представляют собой статичные изображения. Технически документ может содержать файлы любого типа. Их можно просмотреть с помощью сторонней программы, но только изображения видны в самом документе. Это связано с тем, что изображения хранятся в архиве в запакованном виде, а другие файлы в бинарном.

Создадим пустой файл с расширением .docx используя Microsoft Office Word и распакуем его программой 7zip. Содержимое архива отображено на рис. 1.

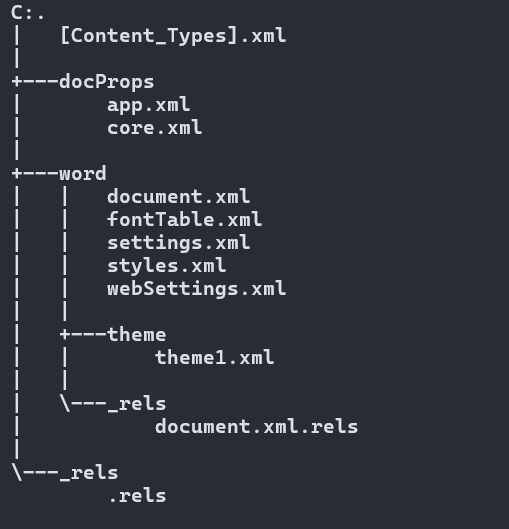


Рисунок 1 – Структура файла .docx

Соглашение об открытой упаковке, разработанное Microsoft, описывает это содержимое следующим образом:

* Content\_Types.xml. Этот файл используется для определения типов содержимого и связанных с ними расширений файлов внутри документа. Эта информация позволяет программам, обрабатывающим файлы .docx, правильно интерпретировать и обрабатывать содержимое документа, обеспечивая его корректное отображение и функциональность при работе с ним;
* docProps/app.xml. Общая информация о документе: количество страниц, слов, символов, название приложения в котором был создан документ и т.п;
* docProps/core.xml. Содержит метаданные, такие как автор документа, описание, дата создания, дата изменения и т.п;
* word/document.xml. В файле хранится вся информация, которую мы видим, открыв документ в редакторе или в читателе. Он содержит не только текст, но и его форматирование;
* word/fontTable.xml. В этом файле перечислены все используемые шрифты в документе;
* word/settings.xml содержит настройки, относящиеся к текущему документу;
* word/styles.xml определяет стили;
* word/webSettings.xml содержит настройки, связанные с веб-публикацией или предназначенные для отображения документа в веб-браузерах;
* word/theme/theme1.xml сосредотачивается исключительно на определении набора цветов, шрифтов и эффектов, который является частью данной темы оформления. Его отличия от других xml файлов заключаются в том, что он специализируется на определении темы оформления, включая цвета, шрифты и эффекты, которые не всегда включены в общие стили. Он упрощает процесс управления оформлением документа, поскольку он позволяет применять сразу целый набор стилей и эффектов, в то время как стили (в styles.xml) могут быть более детализированными и применяться к отдельным элементам. Так же с его помощью к стилям можно подключить элементы, не содержащиеся в других файлах;
* word/\_rels/document.xml.rels обеспечивает связь с внешними ресурсами, например с изображениями или шрифтами;
* \_rels/.rels обеспечивает связь между частями документа.

Документ не ограничивается этими файлами, при добавлении информации его содержимое, как архива, будет меняться. Продемонстрируем это на примере: заполним созданный ранее документ некоторым содержимым. Добавим текст, формулу, изображение. Содержимое файла покажем на рис. 2, а структуру документа на рис. 3.

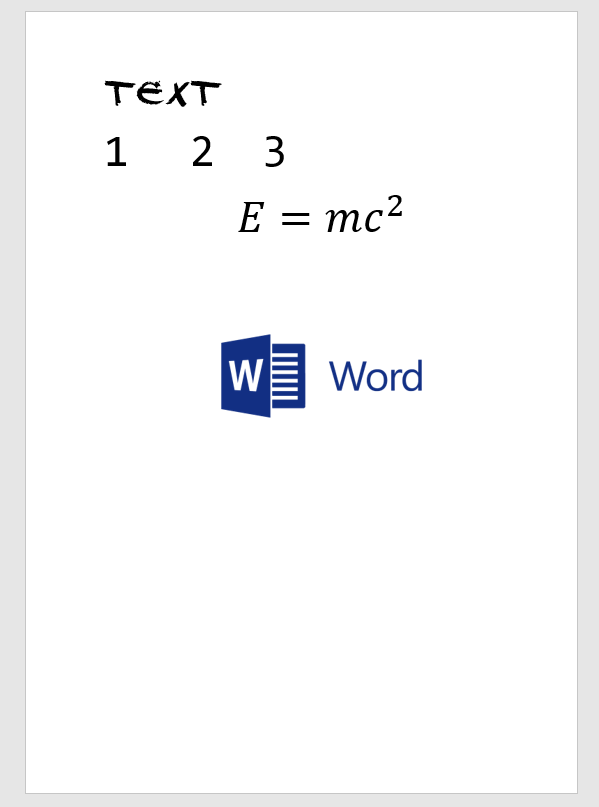


Рисунок 2 – Модифицированный файл .docx

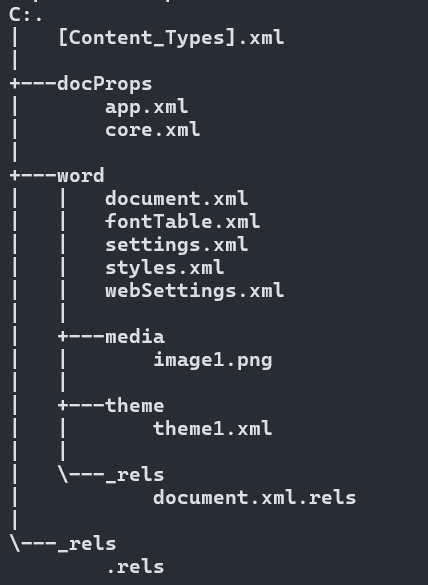


Рисунок 3 – Изменённая структура файла .docx

В папке word появилась папка media, и в ней файл image1.png. Рассмотрим содержимое файла document.xml. Часть документа, содержащая слово «Text», выглядит следующим образом:

…

<w:p w14:paraId="4F9C2915" w14:textId="1D79AD1B" w:rsidR="003F4152" w:rsidRPr="00DA5F0A" w:rsidRDefault="00DA5F0A">

<w:pPr>

<w:rPr>

<w:rFonts w:ascii="DJ Gross" w:hAnsi="DJ Gross" w:cs="Calibri" />

<w:sz w:val="72" />

<w:szCs w:val="72" />

<w:lang w:val="en-US" />

</w:rPr>

</w:pPr>

<w:r w:rsidRPr="00DA5F0A">

<w:rPr>

<w:rFonts w:ascii="DJ Gross" w:hAnsi="DJ Gross" w:cs="Calibri" />

<w:sz w:val="72" />

<w:szCs w:val="72" />

<w:lang w:val="en-US" />

</w:rPr>

<w:t>Text</w:t>

</w:r>

</w:p>

Далее идет часть, содержащая набор цифр, написанных через знак табуляции:

<w:p w14:paraId="5FF2B2BB" w14:textId="69D6CEF1" w:rsidR="00DA5F0A" w:rsidRPr="003C655E"

w:rsidRDefault="00DA5F0A" w:rsidP="003C655E">

<w:pPr>

<w:tabs>

<w:tab w:val="left" w:pos="1843" />

<w:tab w:val="left" w:pos="3402" />

</w:tabs>

<w:rPr>

<w:rFonts w:ascii="Consolas" w:hAnsi="Consolas" w:cs="Calibri" />

<w:sz w:val="96" />

<w:szCs w:val="96" />

<w:lang w:val="en-US" />

</w:rPr>

</w:pPr>

<w:r w:rsidRPr="003C655E">

<w:rPr>

<w:rFonts w:ascii="Consolas" w:hAnsi="Consolas" w:cs="Calibri" />

<w:sz w:val="96" />

<w:szCs w:val="96" />

<w:lang w:val="en-US" />

</w:rPr>

<w:t>1</w:t>

</w:r>

…

Можем заметить, что сначала указываются позиции табуляции, и только потом размещаются цифры на позиции. Абзац, содержащий формулу, в xml имеет больше 50 строк. Отобразим лишь начало данного фрагмента:

<m:oMathPara>

<m:oMath>

<m:r>

<w:rPr>

<w:rFonts w:ascii="Cambria Math" w:hAnsi="Cambria Math" w:cs="Courier New" />

<w:sz w:val="96" />

<w:szCs w:val="96" />

<w:lang w:val="en-US" />

</w:rPr>

<m:t>E=m</m:t>

</m:r>

…

Изображение представлено в документе в виде ссылки <a:blip r:embed="rId4">, которая в файле word/\_rels/document.xml.rels указывает на расположение изображения в архиве:

…

<Relationship Id="rId4" Type="http://schemas.openxmlformats.org/officeDocument/2006/relationships/image" Target="media/image1.png"/></Relationships>

СПИСОК ИСПОЛЬЗОВАНОЙ ЛИТЕРАТУРЫ

1. EPUB 3.3: сайт. – URL: https://www.w3.org/publishing/epub3/ (дата обращения 23.04.2024)
2. EPUB vulnerabilities: Electronic reading systems riddled with browser-like flaws: сайт. – URL: https://portswigger.net/daily-swig/epub-vulnerabilities-electronic-reading-systems-riddled-with-browser-like-flaws (дата обращения 23.04.2024)
3. ECMA-376 ECMA International: сайт. – URL: https://ecma-international.org/publications-and-standards/standards/ecma-376/ (дата обращения 23.04.2024)
4. Gribuser/FB3: сайт. – URL: https://github.com/gribuser/FB3 (дата обращения 23.04.2024)
5. Litres/FB3Reader: сайт. – URL: https://github.com/Litres/FB3Reader (дата обращения 23.04.2024)
6. Litres/FB3Editor: сайт. – URL: https://github.com/Litres/FB3Editor (дата обращения 23.04.2024)
7. Электронные книги и их форматы: DjVu – его история, плюсы, минусы и особенности: сайт. – URL: https://habr.com/ru/companies/maccentre/articles/411545/ (дата обращения 23.04.2024)
8. Леон Ботту; Патрик Хаффнер; Пол Г. Ховард; Патрис Симар; Йошуа Бенджио; Янн Ле Кун (1998).«Высококачественное сжатие изображений документов с помощью DjVu, 7 (3): 410–425»(PDF).Журнал электронных изображений.
9. <https://www.dkriesel.com/en/blog/2013/0802_xerox-workcentres_are_switching_written_numbers_when_scanning>
10. <https://archive.org/post/1053214/djvu-files-for-new-uploads>

План

1. Подготовительный этап:
2. Описание объектной модели
3. Формализация правил
4. Демонстрация работы