**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 2](#_Toc164802418)

[1 ИЗУЧЕНИЕ И ОПИСАНИЕ СУЩЕСТВУЮЩИХ СПОСОБОВ ПРЕДСТАВЛЕНИЯ ДОКУМЕНТОВ И ЛИТЕРАТУРЫ 4](#_Toc164802419)

[1.1 EPub 4](#_Toc164802420)

[1.2 FB2 и его приемник FB3 7](#_Toc164802421)

[1.3 DjVu 9](#_Toc164802422)

[СПИСОК ИСПОЛЬЗОВАНОЙ ЛИТЕРАТУРЫ 10](#_Toc164802423)

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 ИЗУЧЕНИЕ И ОПИСАНИЕ СУЩЕСТВУЮЩИХ СПОСОБОВ ПРЕДСТАВЛЕНИЯ ДОКУМЕНТОВ И ЛИТЕРАТУРЫ

В контексте цифровых данных целесообразно обобщить любой цифровой источник текстовой информации, будь то статья, документ, электронная книга и назвать их одним словом – электронный документ. Такое обобщение 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 которые с ними взаимодействуют создатели или конечные потребители.

Этот раздел будет посвящен наиболее популярным на сегодняшний день форматам хранения и отображения электронных документов, проанализированы преимущества и недостатки данных форматов.

Большинство из них являются контейнерами – представляют собой ZIP архив, хранящий основной текст, как правило в виде XML файла, таблицу стилей, медиаданные и тп.

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, содержащий в себе сайт, так как данный формат 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 ПК, в зависимости от размеров дисплея;
* конвертируемость – данный 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 можно осуществлять из содержащей книгу библиотеки;
* поддержка DRM (Digital rights management англ. – Технические средства защиты авторских прав).

Главный его недостаток связан с информационной безопасностью. Системам чтения, реализующие данный стандарт полностью присуще те же уязвимости, что и веб браузерам. В статье автора 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 данной проблемы является согласие или запрет пользователей на выполнение кода JavaScript в средствах чтения, а также ужесточения требований безопасности.

Данный формат 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 и его приемник FB3

FictionBook (англ. – художественная книга) или FeedBook – это формат хранения электронных книг от российских разработчиков, во главе Дмитрия Грибова, 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 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>

Данный формат обладает такими достоинствами как:

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

Однако данный формат не обделен недостатками:

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

Но все же редко когда художественная литература действительно нуждается в этих элементах. Эти недостатки были исправлены в следующей версии формата FB3. Он 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 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 и его приемник, лишенный недостатков FB3, даже в Рунете.

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 компанией AT&T, специализирующейся на телекоммуникациях. Его разработка была мотивированна распространением отсканированных документов в сети.

Время активного развития интернета можно назвать конец 90-х прошлого века. В то время 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 нецелесообразно. Решить данную задачу распознаванием символов на тот момент было невозможно – точность была не высокая.

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)

План

1. Изучение и описание существующих способов представления документов или литературы для конечного пользователя
2. Подготовительный этап:
3. Описание объектной модели
4. Формализация правил
5. Демонстрация работы