Формат EPUB как замена PDF

EPUB format to replace PDF

А. И. Боровинский,
Библиотека ELiS,
Пермь, Россия

Arsen Borovinsky
ELis Library,
Perm, Russia

Формат EPUB прочно обосновался на рынке электронных книг, но пока еще мало распространен в публичных и вузовских электронных библиотеках. В статье описываются основы формата EPUB, его преимущества и недостатки по сравнению с PDF.

EPUB format has found its niche in the digital book market, still not so popular with public or university e-libraries. The author describes EPUB basic functionalities, its advantages and disadvantages as compared to PDF.

Недостатки PDF

PDF является основным форматом хранения книг уже более 20 лет и стал стандартом де-факто для работы библиотек и всей печатной отрасли. Однако появление мобильных устройств поменяло требование к электронной книге в плане адаптации к размеру устройства отображения. PDF не может полностью решить эту задачу т.к. с самого начала является стандартом с фиксированными размерами страницы (fixed layout).

В PDF есть и другой серьезный недостаток – в нем отсутствует семантическая разметка на слова и абзацы, что осложняет извлечение текста в PDF со сложной верстkой и затрудняет конвертацию в мобильные форматы (reflow) с изменением ширины.

Былое преимущество PDF: сохранение форматирования в разных операционных системах и разных программах отображения стало существенным недостатком в условиях распространения смартфонов. Потребовалась замена PDF, удовлетворяющая потребности мобильных читателей.

Мобильные форматы

Форматы DOC,DOCX,RTF мобильными не являются и не могут эффективно использоваться. Формат TXT не позволяет хранить ничего кроме текста с простейшим форматированием. Из способных адаптироваться к мобильным устройствам распространенным форматам остаются EPUB, MOBI и FB2.

Отечественный формат FB2 не может стать основным форматом в виду его локальности (слабой распространенности в мире). Также он содержит минимальные возможности, ориентирован на художественную литературу и имеет другие недостатки.

MOBI является вендор-зависимым форматом для Amazon Kindle.

Методом исключения остался самый распространенный мобильный формат: EPUB.

Основы EPUB

EPUB является открытым стандартизованным[1] форматом для электронных публикаций. Файл EPUB является ZIP-архивом, в котором находятся остальные файлы. Цель упаковки в ZIP не сжатие текста, а размещение в одном контейнере нескольких файлов книги.

Внутри EPUB находится несколько XML-файлов[2] с описанием структуры EPUB, содержания, метаинформации и spine[3] – списка файлов в архиве с указанием их MIME-типов. Само содержимое книги представляет из себя HTML-страницы в XML-представлении (XHTML). XHTML-страницы книги могут включать в себя все возможности HTML, включая выполнение JavaScript, отправку

[1] http://idpf.org/epub
[2] http://www.idpf.org/epub/301/spec/epub-ocf.html
[3] http://www.idpf.org/epub/301/spec/epub-publications.html#sec-spine-elem

онлайн-форм, вставку изображений и мультимедийного содержимого как из самого файла, так и из интернета.

Такая простая структура позволяет легко генерировать или изменять существующие файлы EPUB на основе знаний обычного HTML без использования специализированных библиотек после непродолжительного обучения формированию структуры обязательных XML-файлов.

При работе с EPUB необходимо значительное внимание уделять программам чтения книг. Разработчики формата предполагают, что программы будут изменять отображение книги, например, менять размер шрифта с учетом пользовательских предпочтений. Многие требования спецификации являются опциональными и не обязательными к реализации в конкретной программе чтения.

Разработчики EPUB учли возможность чтения на устройствах с малым объемом оперативной памяти. Для этого книга может быть разбита на произвольное число XHTML страниц чтобы не пришлось загружать большой объем книги в память устройства. В этом смысле классических книжных страниц в EPUB не существует. Цель разбиения на страницы XHTML – экономия памяти и упорядочивание книги для авторов. Большинство программ чтения визуально сливает все XHTML-страницы в одну книгу, а пользователю отображают логическую нумерацию, рассчитывающуюся от размеров экрана и меняющуюся при смене ориентации устройства.

Преимущества EPUB перед PDF

Основные преимущества формата связаны с использованием возможностей HTML, CSS и JavaSript. Можно выделить следующие отличия от PDF:

1) Семантическая разметки, слов и абзацев.
2) Адаптация к экрану пользователя (размерам и плотности пикселей) с помощью CSS media queries[4]. Верстальщик может адаптировать поведение книги к различным мобильным устройствам или для просмотра на ПК.
3) Возможность программным путем изменить размер шрифта или включить ночной режим.
4) Большое число программистов знакомо со стеком технологий HTML/CSS/JavaScript, что дает низкий порог вхождения для разработки собственных книг и позволяет использовать готовые библиотеки для анимации эффектов на JavaScript.
5) EPUB может использоваться в качестве контейнера для длительного хранения веб-страниц с мультимедийным содержимым.

Недостатки EPUB

Программы для чтения EPUB могут значительно различаться в реализации процесса чтения электронных книг. Перелистывание страниц может быть реализовано совершенно по-разному. В одних программах будет составляться логическая нумерация и при ее составлении текст не будет отсекаться на середине слова, в других будет использоваться кинетический скроллинг, в третьих пользователь сможет выбрать любой из упомянутых режимов.

Ссылаться внутри EPUB на определенную отображаемую страницу больше нельзя. Необходимо развитие библиографии для возможности составления библиографической ссылки в EPUB. Сам EPUB стандартизовал машиночитаемый канонический идентификатор фрагментов[5] текста.

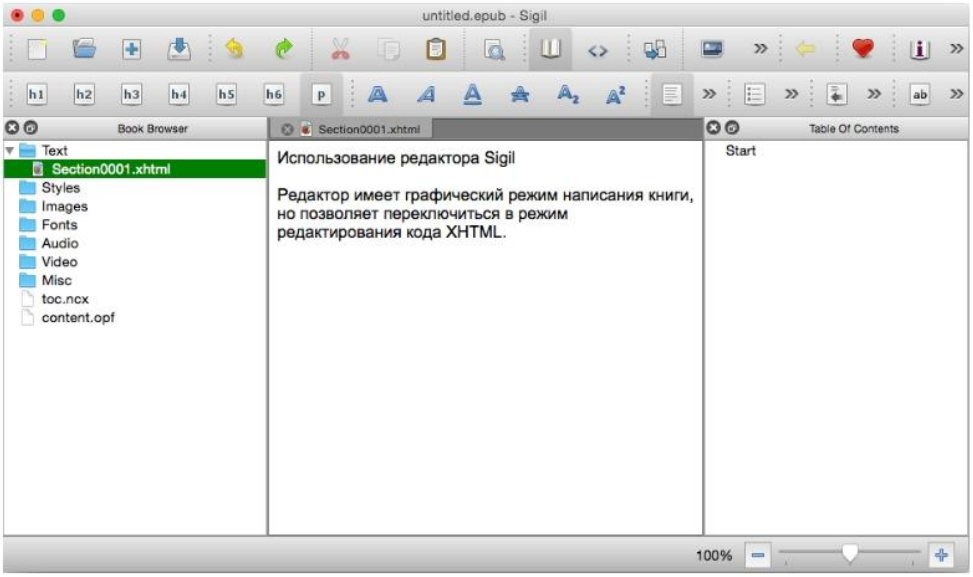
Программы чтения EPUB сильно различаются в полноте реализации спецификации и могут вмешиваться в рендеринг страницы. Из-за этого может искажаться сложная верстка и усложняется задача тестирования на различных устройствах и в приложениях.

Не смотря на длительное время прошедшее с появления EPUB, существует дефицит инструментов для написания книг. Популярный и бесплатный Apple iAuthor работает с EPUB фиксированных размеров, что существенно уменьшает преимущество EPUB по сравнению с PDF. Остаются специализированные программы Blue Griffon и Sigil. Adobe InDesign позволяет экспортировать

[4] http://www.idpf.org/epub/301/spec/epub-mediaoverlays.html
[5] http://www.idpf.org/epub/linking/cfi/epub-cfi.html

книги в EPUB, но после экспорта может потребоваться ручная правка документа или предварительная подготовка экспортируемой книги[6].

Кроме неполного выполнения стандарта, программы чтения EPUB часто нарушают стандарт. Например, большинство современных программ воспроизводят встроенное в EPUB2 видео, хотя в стандарте поддержка видео появилась только в EPUB3.



Sigil – бесплатный EPUB-редактор

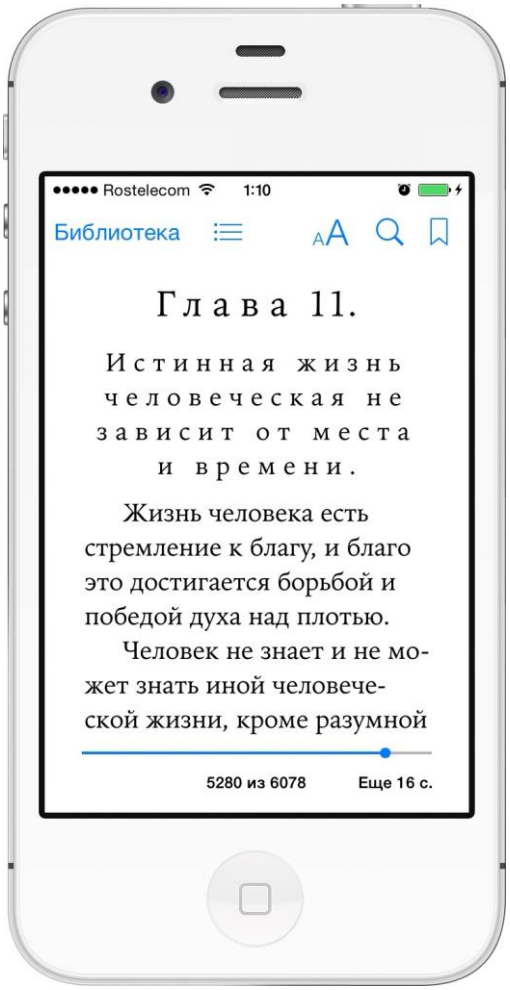
Паритет возможностей

PDF развивается уже длительное время и вобрал в себя часть возможностей веб-страниц. Поэтому заметная часть функционала PDF пересекается с функционалом EPUB. Перечислим одинаковые возможности EPUB и PDF:

1) Возможность встраивания JavaScript. JavaScript может использоваться для некоторых типов атак на пользователей, поэтому часто эта возможность отключена как для PDF, так и для EPUB. Однако, т.к. веб-страница знакома программистам значительно лучше объектов PDF, можно ожидать более полного раскрытия потенциала этой возможности. Многие упомянутые далее возможности EPUB могут потребовать работы JavaScript.
2) Встраивание видео и аудио. Встраивание возможно в обоих форматах, однако, существующие средства разработки для EPUB являются более дружественными.
3) DRM-защита может использоваться в обоих форматах. EPUB не имеет какой-то встроенной общепринятой системы защиты и каждая система защиты может реализовать ее своим способом. В PDF есть защита по паролю (который является ключом для дешифровки текста) и сертификату. Наиболее распространена и является стандартом де-факто DRM-защита компании Adobe.
4) Цифровая подпись стандартизована для обоих форматов.
5) Встроенные формы.
6) Содержание книги. В обоих форматах возможно построение содержания книги. В PDF содержание называется Bookmarks и ссылается на страницы. В EPUB часто используется ссылка на якорь на XHTML-странице. В EPUB содержание описывается XML-документом, но может быть и отдельной XHTML-страницей для придания специфического внешнего вида.

[6] http://www.whyepub.ru/from_in_design_to_epub/

7) Расширяемость. В обоих форматах возможно встраивание собственных объектов. Например, в оба формата можно встроить описание книги в формате MARCXML. Преимуществом EPUB опять является большая простота реализации встраивания.
8) Метаинформация. В PDF используется pdfmark для описания автора, заглавия, даты публикации и ключевых слов. В EPUB для аналогичной задачи применяется[7] схема Dublin Core.



Толстой в EPUB

Толстой в PDF

Варианты использования EPUB

EPUB может использоваться одновременно с PDF-версией книги в процессе издания. Библиотеки должны уметь при этом хранить оба объекта и, желательно, при скачивании предоставлять пользователю выбор в каком формате скачать книгу. Фактическое содержимое книги в формате EPUB может при этом не соответствовать PDF-версии и реализовывать большее количество возможностей. При скачивании на мобильное устройство следует рекомендовать пользователю книгу в формате EPUB как адаптированный формат.

EPUB может использоваться для интерактивных книг, в которых с помощью JavaScript и CSS производится анимация, добавляются игровые элементы, видео и звуковое сопровождение. В

[7] http://www.idpf.org/epub/30/spec/epub30-publications.html#sec-metadata-elem

России данным направлением занимаются компании Орфограф[8] (учебники для школы в приложении «Дай 5!») и Enaza (платформа для создания интерактивных книг UnderPage[9]).

Геймификация в интерактивной книге Принцесса на горошине.
Требуется расставить портреты

Современные СМИ переходят в онлайн и печатные выпуски газет не отражают мультимедийного содержимого встречающегося на сайте. EPUB может служить контейнером для архивации и распространения выпусков электронных газет сайтов современных СМИ. Возможно EPUB не является оптимальным форматом для долгосрочного хранения (через некоторое время ссылки на внешние ресурсы станут неработоспособны и требуется механизм работы ссылок между статьями одного издания), но другой альтернативой является создание зеркала сайта, с которым тоже есть ряд проблем.

В основном презентации создают в PowerPoint в формате PPT/PPTX. Для длительного хранения этот формат не оптимален из-за зависимости от Microsoft Office и презентации часто сохраняются в PDF. Однако, в PDF презентации не имеют анимации и ряд функций социального продвижения. В качестве варианта, презентации из формата PDF можно конвертировать в EPUB с добавлением недостающего функционала. Использование EPUB в части его интерактивных возможностей позволяет создавать и передавать на длительное хранение презентации в формате близком к современным онлайн-сервисам, таким как SlideShare, не опасаясь при этом привязки к одному производителю.

Научные статьи долгое время ориентировались на PDF для печати результатов и Microsoft PowerPoint для презентации на конференциях. Выбор PowerPoint при этом продиктован необходимостью встраивания видео, наглядно демонстрирующего результаты исследования. EPUB позволяет реализовать интерактивные возможности в статье и, за счет встраивания JavaScript, предоставляет платформу для создания качественно новых научных статей, в которых результаты исследований представлены в виде масштабируемых графиков с возможностью изменения шкалы, всплывающих подсказок и анимацией. Конечно, имеется возможность встраивания видео и форм обратной связи с авторами. Для научных статей естественнонаучных специальностей важно отображение формул. В EPUB имеется возможность использовать MathML, однако не все про-

[8] http://orfogr.ru/
[9] http://underpage.com/

граммы для чтения его поддерживают. В случае отсутствия поддержки имеется возможность встроить JavaScript-библиотеку MathJax[10] для вывода формул средствами CSS.

Карты с самого начала плохо встраиваются в PDF т.к. часто требуют возможности масштабирования. Распространение карт в виде рисунков высокого разрешения тоже не является приемлемым вариантом т.к. нет уверенности в возможностях удобного просмотра карты со стороны пользовательского ПО. Выходом из этой ситуации с самого начала были специализированные веб-плееры, позволяющие удобное масштабирование на сайте. Такие плееры могут быть перенесены в EPUB вместе с необходимой картой или сборником карт.

Комиксы, являясь простыми иллюстрациями, легко могут быть перенесены в формат EPUB. В зависимости от размеров экрана, можно формировать вывод комиксов с разным числом колонок.

Оцифрованные книги занимают значительное место в современных библиотеках, а одна цифровая копия может занимать сотни мегабайт. Формат EPUB может заменить в этой задаче PDF с тем ограничением, что оцифрованные страницы представляют из себя рисунки и не могут быть автоматически адаптированы для чтения с малых экранов, т.е. отсутствует главное преимущество EPUB перед PDF. Тем не менее, положительным преимуществом является возможность извлечения оцифрованных страниц простой распаковкой EPUB-файла как ZIP-архива. Возможность делать текстовые подкладки из распознанного текста сохраняется.

Различия EPUB2 и EPUB3

Фактическое хождение в сети получили вторая и третья версия спецификации EPUB. Основным улучшением EPUB3 стала поддержка HTML5 (XHTML5) и связанные с ними CSS3, видео и аудио. EPUB2 основан на XHTML4 и, согласно спецификации, не поддерживает видео. Часть программ расширяют EPUB2 поддержкой элемента Video и называют такой формат EPUB2.5, хотя официально EPUB2.5 никогда не существовало, а развитие EPUB2 прекращено.

С появлением EPUB3 была переработана структура служебных XML-файлов с возможностью сохранить обратную совместимость с программами для чтения EPUB2 путем дублирования навигации.

Большинство книг в EPUB подготовлены в формате EPUB2, однако, в виду прекращения его развития, будущие книги целесообразно выпускать в EPUB3.

С точки зрения технической реализации, в виду высокой сложности HTML5, программы для чтения EPUB3 строятся на основе браузеров. XHTML4 – более простой формат и имеет большое число готовых библиотек, что позволило создать программы для рендеринга EPUB2 без использования браузера. Отказ от браузера позволяет более эффективно управлять энергопотреблением устройства, но несет в себе потенциальные ошибки в реализации уже не только EPUB, но и XHTML.

Заключение

С точки зрения перспективы формата, в ближайшее время PDF и EPUB будут сосуществовать и появление конкурирующего мобильного формата маловероятно. По мере замены мобильными устройствами персональных компьютеров, EPUB будет вытеснять PDF.

EPUB нельзя назвать стабильным форматом т.к. его развитие связано с активным изменением веб-технологий и, при переходе на новую версию, может произойти отказ от части функционала, являющегося расширением EPUB над XHTML. В тоже время, разработчики браузеров и веб-стандартов тщательно заботятся об обратной совместимости и в среднесрочной перспективе можно ожидать совместимость браузеров с используемом в EPUB кодом XHTML.

Несмотря на трудности с форматом EPUB, его распространение через магазины электронных книг уже активно идет. Библиотеки и издатели, желающие остаться актуальными для современных читателей, тоже должны освоить новый формат.

[10] https://www.mathjax.org/