Широков Александр

ИЗГОТОВЛЕНИЕ ЭЛЕКТРОННЫХ КНИГ. СОВЕТЫ И РЕКОМЕНДАЦИИ

ПРАКТИЧЕСКОЕ ПОСОБИЕ

(редакция от 11.10.2022)

Кто знает аз да буки, тому и книги в руки
(пословица)

Создание электронных версий некогда изданных книг и журналов – занятие, которое трудно назвать простым, к тому же сложившаяся в России ситуация такова, что дело это, за немногими исключениями, в основном лежит на плечах добровольцев-энтузиастов, армия которых хоть и многочисленна (ну что ж поделаешь – не перевелись ещё у нас в стране истинные библиофилы, несмотря ни на чьи старания!), но весьма разрознена. Иными словами, по-настоящему централизованно и профессионально сохранением в электронном виде копий подчас бесценных книжных изданий, являющихся неотъемлемой частью нашего культурного наследия, мало кто занимается.

По этой причине не существует общепринятых требований и правил, которые должны предъявляться к электронным книгам и быть обязательными к выполнению при их создании. Ввиду этого изложенные в настоящем пособии советы по созданию электронных книг выражают лишь собственные соображения автора по этому вопросу, и, как следствие, носят исключительно рекомендательный характер.

Содержание:

От автора

Данное пособие было написано с целью поделиться опытом в области создания электронных книг, а заодно попытаться по-своему систематизировать и обобщить имеющуюся информацию по этой теме. Я постарался изложить всё в виде совокупности практических советов. При этом мне пришлось учитывать следующее.

Во-первых, в оцифровке печатных изданий есть много разнообразных нюансов и тонкостей, иногда взаимно обусловленных и связанных друг с другом, что осложняет последовательное и логичное изложение имеющихся по тексту объяснений. В связи с этим основные усилия как раз и были направлены на разрешение этой проблемы – чтобы общий ход подачи материала выглядел в достаточной мере стройно.

Во-вторых, при создании электронных книг нужны навыки владения некоторым количеством прикладных компьютерных программ, также весьма желательна и кое-какая теоретическая подготовка. Именно поэтому по тексту по возможности приводятся ссылки на различные внешние источники, рекомендуемые для самостоятельного изучения и содержащие более полную и подробную информацию по тематике, непосредственно относящейся к контексту.

В-третьих, оцифровка книг – ремесло относительно молодое, не имеющее твёрдо устоявшихся традиций, в связи с чем мнения по одним и тем же вопросам часто бывают диаметрально противоположными. Многие утверждения в пособии могут быть оспорены, поэтому не стоит воспринимать его содержимое как сборник неких правил, требующих неотступного выполнения – на этот счёт есть очень хороший афоризм: «Инструкции призваны заменить человеку голову, а голова – инструкции».

На данный момент это пятая версия пособия (предыдущие редакции были датированы 29.08.2010, 05.12.2010, 05.03.2011 и 30.12.2011) и не исключено, что оно и в дальнейшем будет редактироваться и дополняться. Самую свежую его версию всегда можно скачать с моего сайта. Пособие является моей интеллектуальной собственностью, однако будучи его автором я позиционирую его как свободно распространяемое произведение. Под этим понимается, что любой желающий имеет право бесплатно загрузить себе копию настоящего пособия (файл в формате CHM или PDF) для личного пользования, либо разместить его у себя на сайте при условии указания автора и ссылки на родной сайт. Я не разрешаю коммерческое распространение пособия, а также самовольное изменение имеющейся в его chm- или pdf-файле информации и распространение в таком модифицированном виде без моего ведома.

Форматы книг

Электронные книги, с которыми сейчас могут иметь дело пользователи, бывают представлены в виде файлов самых различных типов. В настоящем пособии акцент будет сделан на следующих трёх.

DJVU – формат, созданный специально для размещения в сети Интернет отсканированных с хорошим разрешением изображений, содержащих текстово-графическую информацию. Изображение в этом формате, как правило, состоит из трёх слоёв: первый – “background”, содержит фон и малоконтрастные участки с плавными переходами оттенков, второй слой – “mask”, чёрно-белый трафарет-маска с высококонтрастными участками, такими как текст, схемы, графики, и третий слой – “foreground” с информацией о цвете для слоя маски. Слои “background” и “foreground” сжимаются вейвлет-алгоритмом IW44, а слой “mask” упаковывается методом JB2. В результате этого достигается очень высокая степень компрессии, благодаря чему DJVU очень удачно сочетает в себе относительно малые размеры выходного файла и довольно неплохое качество отсканированного изображения/текста, несмотря на то, что алгоритмы IW44 и JB2 являются методами сжатия с потерями. Формат сейчас усовершенствован – в него добавлен ещё и слой распознанного (OCR) текста. В целом, DJVU оказался особенно хорош для перевода в электронный вид старых книг, справочников и учебников, изданных в «докомпьютерную эру».

Перечень программ, позволяющих осуществлять просмотр файлов в таком формате, довольно обширен и здесь будет вполне достаточным назвать “Djvu Solo” и “WinDjView”.

PDF – формат, основанный на языке программирования PostScript и облюбованный типографами за то, что позволяет подготавливать в электронном виде разнообразные публикации, обеспечивая при этом полное соответствие между напечатанной страницей и её отображением на экране компьютера. Формат использует для данных алгоритм сжатия без потерь, что позволяет получать относительно малого размера файлы многостраничных книг, содержащих большие объёмы текстовой информации и широко применяется для создания документации технического характера, а также электронных версий научных статей.

Язык PostScript предназначен для описания графических объектов, характеристики которых задаются при помощи математических уравнений, то есть он, прежде всего, нацелен на работу с векторной графикой (к ней относятся и изображения символов шрифтов типа True Type). В силу этого формат PDF тоже можно считать векторным, однако в него могут быть

внедрены и растровые (фотографические) изображения. Благодаря этой особенности встречаются pdf-книги, которые представляют собой набор отсканированных страниц бумажного издания, и в этом плане они в чём-то подобны книгам формата DJVU.

Традиционно для просмотра pdf-файлов используется программа “Adobe Acrobat”. Сейчас появилось много альтернативных программ, также умеющих открывать файлы рассматриваемого типа, среди них – популярные браузеры “Mozilla Firefox” и “Google Chrome”.

CHM – изначально задумывался как формат файлов справки для операционных систем (ОС) семейства Windows. Представляет из себя набор веб-страниц, сопутствующих им рисунков GIF, JPEG или PNG и других вспомогательных файлов (например, таблиц стилей), упакованных (скомпилированных) в один файл. Текстовая информация в нём подвергается сжатию, чем достигается сравнительно небольшой размер конечного файла. Формат вполне популярен, поскольку позволяет создавать электронные книги с красивым оформлением и удобной навигацией. Кроме того, данный формат используется ещё и для создания офф-лайн версий интернет-сайтов.

Файлы в формате CHM хороши прежде всего тем, что для их просмотра в среде “Windows” (начиная с 98-го) не требуется установки дополнительного программного обеспечения – они по умолчанию открываются системной утилитой “hh.exe”. Вероятно по этой причине количество сторонних программ-читалок файлов этого формата невелико, но, тем не менее, они есть, например, это “Ice Book Reader Professional” (“ice-book.ru”).

Кроме перечисленных типов файлов, электронные книги могут быть представлены в форматах TXT, HTML, RTF, DOC, FB2 и некоторых других. Для указанных файлов можно применять программу “Cool Reader”. Кратко о рассматриваемых форматах:

- TXT – формат «только текст», исторически самый старый. Данные файлы – наиболее универсальные, информация из них может быть прочитана самыми разными приложениями, работающими в различных операционных системах. Главное достоинство – небольшой размер файлов, а недостаток – невозможность сохранить форматирование текста.
- HTML – формат веб-страниц и, соответственно, основная сфера его применения – Интернет. Файлы этого формата достаточно компактны и при этом помимо самого текста содержат в себе информацию о его форматировании. Содержимое файла в таком формате можно просмотреть не только при помощи любого текстового редактора, но и посредством более специализированных программ – браузеров.

- RTF – аббревиатура, являющаяся одновременно и расширением файлов такого типа, расшифровывается как “Rich Text Format” – «Расширенный текстовый формат». В соответствии с таким названием, rtf-файл способен хранить в себе самое разнообразное форматирование содержащегося в нём текста. Основной недостаток, присущий файлам этого типа – сравнительно большой размер.
- DOC, DOCX – собственный формат документов Microsoft (далее – MS) “Word” (DOCX – в MS Word 2007-2021, DOC – в более ранних версиях этого текстового редактора). По причине огромной популярности данной программы де-факто стал ещё одним форматом для электронных книг. При отсутствии установленного на компьютере «родного» приложения из пакета MS Office doc/docx-файлы можно открыть и просмотреть редактором “Writer” из пакета LibreOffice (далее – LO).
- FB2 – файлы формата “Fiction Book”, основанного на языке XML. Формат разработан специально для создания электронных книг с определённой структурой хранящейся в ней информации, что облегчает автоматическую обработку таких книг при размещении их в интернет-библиотеках.

Общие положения

Так как книги – это, прежде всего, продукт творческого труда, то разговор об их оцифровке нужно начать с напоминания о том, что на территории нашей страны вопросы, касающиеся результатов интеллектуальной деятельности, регламентируются частью четвёртой Гражданского Кодекса России. Применительно к нашей теме необходимо уяснить для себя следующее: законно изготавливать электронную копию книги можно только при наличии соответствующего разрешения обладателя авторских прав на неё, а в случае отсутствия такого разрешения не запрещается оцифровка лишь тех книг, которые перешли в общественное достояние в соответствии со ст. 1282 ГК РФ.

При оцифровке печатных изданий далеко не последнюю роль играет комплекс технических характеристик компьютера, при помощи которого будет изготавливаться электронная книга – объём оперативной памяти, тактовая частота процессора и прочее. В целом, чем выше эти параметры – тем лучше, ведь тогда компьютер способен быстрее осуществлять обработку информации, преобразуемое количество которой при создании электронной книги обычно очень велико. В данном пособии в процессе изложения материала будут упоминаться программные средства, работающие под управлением операционных систем семейства Windows, поскольку разные их версии наиболее распространены у пользователей. Приверженцам же Linux, зачастую являющимся людьми достаточно опытными в информационных технологиях, не составит большого труда адаптировать информацию из пособия под предпочитаемую ОС.

Зададимся вопросом: если, по большому счёту, не существует никаких общепринятых стандартов и требований по оформлению электронных книг, то не пришла ли пора для тех, кто занимается их изготовлением, определить хоть какие-нибудь правила, чтобы руководствоваться ими? Я предлагаю к использованию два следующих принципа, кратко которые можно сформулировать так:

- файл электронной книги должен иметь по возможности минимальный размер
- в электронном варианте книги недопустимо существенное искажение информации по сравнению с её бумажным оригиналом

Приведённые тезисы нуждаются в развёрнутом комментарии. Несмотря на то, что компьютерная техника развивается бешеными темпами – непрерывно растут ёмкости носителей информации, быстродействие процессоров и пропускная способность интернет-соединений – считаю, что всё-таки лучше, когда электронная книга имеет в байтовом выражении по возможности меньшую величину. В этом случае на одном и том же носителе можно записать гораздо больше информации, так как экономится место, к тому же проще осуществить пересылку такой книги по сети. Именно по этой причине в настоящем пособии внимание будет уделено созданию книг форматов DJVU, PDF и CHM как максимально полно удовлетворяющих первому принципу. Относительно PDF сразу следует сделать дополнительную оговорку. В пункте «Форматы книг» уже упоминалось, что встречаются pdf-книги являющиеся набором отсканированных страниц бумажного издания. Такой файл по своему размеру почти всегда проигрывает (и порой весьма значительно) аналогичному файлу формата DJVU, не имея при этом сколь-нибудь заметного преимущества в качестве изображения. Отсюда следует, что книги в формате PDF с цельными сканами страниц также не удовлетворяют принципу минимального размера файла, по коей причине создание именно таких pdf-книг в данном пособии рассматриваться не будет.

При попытках минимизации файла не следует перегибать палку. Мне встречались djvu-книги, страницы которых были не очень хорошо отсканированы (с довольно низким разрешением – для уменьшения размера сканов страниц), а из-за того, что при создании djvu-файла изображение и так подвергается сжатию с потерями информации, текст на страницах таких электронных книг получился трудночитаемым. Это подводит нас к рассмотрению второго принципа – отсутствия в оцифрованном издании существенного искажения информации. Электронная копия книги должна наиболее полно соответствовать тому, что она – именно копия, максимально точно дублирующая имеющееся в бумажном оригинале его информационное наполнение, ведь только в этом случае данными из такой книги можно пользоваться уверенно, не опасаясь сделать непреднамеренную ошибку и имея возможность, например, с чистой совестью сослаться на такой литературный источник также, как и на

печатный. Согласитесь, что навряд ли вызовет большое доверие электронный вариант бумажного издания, если, допустим, в нём будет изобилие опечаток, обусловленных тем, что сканированные страницы прогнали через программу распознавания текста и не потрудились как следует проверить результат её работы – ничего, кроме раздражения и недоумения, такая оцифрованная книга у читателя не вызовет.

Считаю допустимым, что в особых, исключительных, случаях в электронную книгу можно вносить некоторые корректировки её содержимого, но они, во-первых, непременно должны быть обоснованы, во-вторых, делать это нужно деликатно, а, в-третьих, будущего читателя такой книги нужно обязательно известить о том, какому именно редактированию подверглась исходная информация.

Дальнейшее изложение материала в пособии подчинено следованию реализации на практике двух рассматриваемых принципов. Строго говоря, положения о минимальном размере файла и об отсутствии искажений информации могут вступать в противоречие друг с другом. Здесь очень важно умение найти оптимальное решение и весьма существенным тут является выбор формата, в котором лучше будет сохранить оцифрованную книгу, а это зависит от её содержимого – я могу лишь посоветовать ориентироваться на составленную мной Таблицу 1, которая приводится ниже.

Таблица 1. Некоторые типы информационного наполнения книг и предпочтительные для них форматы

Издание	**Формат**
Книги при изготовлении электронных версий которых требуется сохранить композицию и внешний вид их страниц (старинные и раритетные издания)	DJVU
Книги, содержащие текст без иллюстраций (произведения художественной литературы, публицистические работы и т. п.) при условии, что требуется сохранение именно текстовой информации, а также книги с малым или умеренным количеством иллюстративного материала	CHM, PDF
Литература научного и технического характера, имеющая на своих страницах текст со значительным числом специальных знаков и формул, рисунков, схем, а также иные книги с обилием иллюстративного материала	DJVU
Периодические издания – журналы, газеты	DJVU

Процесс создания электронной книги является многоэтапным. В самом общем виде ход её изготовления может быть представлен как последовательность следующих стадий:

1. Сканирование – получение цифровых копий страниц книги (издания) в виде совокупности графических файлов (сканов);

2. Обработка сканов – редактирование и корректировка полученных изображений, прежде всего с целью улучшения их качества;

3. Распознавание сканов – программный перевод имеющихся в графических файлах текстовых блоков в компьютерный текст;

4. Корректура – вычитка распознанного текста и исправление ошибок в нём;

5. Подготовка иллюстративного материала – создание на основе сканов файлов рисунков, схем, диаграмм, графиков отдельно от основного текста, а также подготовка таблиц и проверка содержащихся в них данных;

6. Вёрстка – форматирование текста и сведение его воедино с подготовленным иллюстративным материалом, а также создание общего оформления электронной книги;

7. Создание выходного файла – запись в выбранном формате готового файла-книги.

Наличие или отсутствие в приведённой схеме каких-либо конкретных стадий также находится в зависимости от содержания книги и от того, в каком именно формате предполагается создание её цифровой копии. Далее каждый из перечисленных этапов формирования электронной книги будет рассмотрен более детально.

Сканирование

Эта стадия является обязательной и её по праву можно назвать самой ответственной, поскольку именно от неё зависит качество будущей электронной книги.

Прежде всего, нужен сканер подходящего типа. Спектр выпускающихся разновидностей этих устройств очень широк. Существуют модели, у которых и передача данных, и их электропитание осуществляются через какой-либо компьютерный порт (как правило, USB), то есть через него происходит подача энергии, необходимой для питания лампы сканера и работы перемещающего её электродвигателя. Это не позволяет устанавливать в такие сканеры двигатели большой мощности, по причине чего скорость поступательного движения каретки и, соответственно, скорость сканирования страницы, оказывается довольно низкой. Указанное обстоятельство весьма критично для сканирования многостраничных книг, ведь зачастую именно оно является ещё и «лимитирующей стадией» всего процесса в целом. Сканеры описанного типа хороши лишь в случаях, когда нужно отсканировать порядка десяти страниц, но не более того. Поэтому рекомендую выбирать сканер, у которого запитка его механических компонент осуществляется от обычной электрической сети (пусть даже и через адаптер).

Второй важный момент, на который следует обратить внимание – это программное обеспечение. При покупке сканера в комплект его поставки входят не только необходимые драйвера, но и программа для управления режимами работы, приёма данных с этого устройства, первичной их обработки и сохранения в виде файлов.

К сожалению, не все подобные программы хорошо подходят для работы с многостраничными книгами. Дело в том, что в некоторых из них для получения одного скана нужно каждый раз выполнять пусть и короткую, но всё же определённую последовательность действий, сводящуюся к одним и тем же операциям: выбор пунктов меню, нажатие кнопок в диалоговых окнах, установка опций, что осуществляется посредством нескольких щелчков

мышью. При сканировании большого количества страниц это также заметно тормозит весь процесс, как и описанные выше медлительные сканеры, но вдобавок ещё и может сильно раздражать.

По указанной причине советую пользоваться программой, аналогичной “OmniPage”. В ней сперва выставляются необходимые параметры сканирования, а затем подаётся команда к его началу. После того, как страница будет отсканирована, программа каждый раз выдаёт маленькое диалоговое окно (Рисунок 1) с двумя кнопками, первая из которых (“Stop loading pages” – «прекратить загрузку страниц») позволяет остановиться, а чтобы продолжить сканирование дальше, нужно приложить к стеклу сканера следующую страницу и нажать на вторую кнопку “Add more pages” («добавить ещё страницы»). Таким образом “OmniPage” позволяет сначала получить группу сканов, просмотреть их, в случае необходимости отбраковать неудачные и лишь потом записать в виде совокупности файлов.

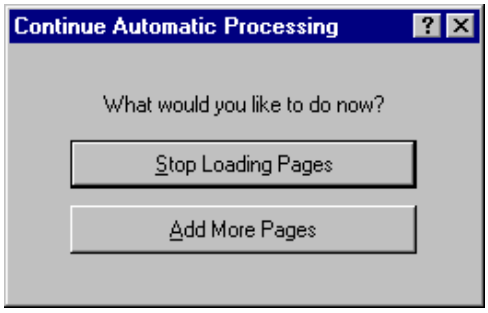


Рисунок 1. Вид диалогового окна в программе “OmniPage” после сканирования страницы

Теперь следует рассказать о параметрах сканирования, ведь именно они определяют качество скана. Прежде всего это – разрешение изображения, измеряющееся в dpi (dots per inch – количество точек-пикселей на дюйм). Рекомендуемое значение, к тому же и наиболее часто применяемое – 300 dpi. Использовать более низкое можно, но далеко не всегда целесообразно. В случае, если текст сканируемых страниц набран мелким шрифтом, то имеет смысл проводить сканирование с более высоким, нежели 300 dpi, разрешением (400 или 600 dpi), если только возможности самого сканера позволяют это делать. При этом необходимо помнить, что чем выше разрешение изображения, тем больше (при прочих равных условиях), будет размер файла-скана.

Не менее существенное значение имеет и рациональный выбор режима сканирования. Три основные их них следующие: цветной, оттенки серого и чёрно-белый.

Цветной (“color”) режим – это получение картинки (скана) с глубиной цвета (по-английски – “color depth”) в 24 бит, то есть изображения со всеми возможными (точнее – заметными среднему человеческому глазу) переходами оттенков, которых в этом случае можно закодировать аж $2^{24} = 16\,777\,216$ штук. Данный режим сканирования наиболее подходит для страниц, содержащих цветные иллюстрации или разноцветные надписи (например, заголовки в некоторых изданиях могут иметь цвет, отличный от цвета основного текста). Также он может

понадобиться в случае потребности сохранить не только сам текст, но и внешний вид его носителя – фактуру и цвет бумаги, какие-нибудь мелкие дефекты типа царапин, чьих-то рукописных пометок, обтрепавшихся от времени уголков страниц и т. п.

При сканировании в режиме оттенков серого, который часто в англоязычных программах обозначается словом “grayscale” («шкала серого») – это получение скана с глубиной цвета 8 бит (с гаммой из $2^8 = 256$ оттенков серого цвета, включая чёрный и белый как крайние). Если выражаться простым житейским языком, то визуально такой скан выглядит как чёрно-белая фотография. Этот режим сканирования хорош для страниц с чёрно-белыми иллюстрациями, содержащих различные оттенки серого цвета. При сканировании в этом режиме также можно сохранить, правда, в нецветном варианте, внешний вид носителя текста – например, всё ту же фактуру бумаги.

Третий режим сканирования называется чёрно-белым (“black and white”), ещё он может именоваться одноцветным или монохромным (“monochrome”). Его не следует путать с описанным выше режимом оттенков серого. В чёрно-белом режиме получаемый скан имеет глубину цвета 1 бит. По сути это означает, что при сканировании светлые участки страницы воспринимаются только как имеющие белый цвет, а тёмные – только как чёрный. Наиболее близкая жизненная аналогия такого режима – это ксерокопия, на которой зачастую мелкие или слабовыраженные детали страницы (ну например, слишком блеклый оттиск печати на документе) не фиксируются. Чёрно-белый режим наиболее подходящ для страниц, содержащих текст, таблицы и рисунки, представляющие собой совокупности контрастных, чётко очерченных линий (например, графики и схемы), в случаях, когда нужно запечатлеть именно информационное содержание страницы, отметя фон.

Некоторые программы управления сканером позволяют получать изображения в других режимах, отличных от рассмотренных трёх основных. Например, это могут быть варианты сканирования в 16- или 256-цветной палитре. Лично у меня ни разу не возникало необходимости в подобных режимах, даже при наличии возможности их использования, хотя представляется, что нельзя исключать ситуацию, когда их применение будет вполне оправданным.

Большинство программ для работы со сканером позволяют автоматически выполнять некоторые корректировки сканов-изображений, например, задавать предустановки для контрастности, яркости, цветового баланса. Зачастую это позволяет получать сканы заметно более высокого качества, что в дальнейшем упрощает дальнейшую их обработку, поэтому перед сканированием можно рекомендовать немного поварьировать этими настройками и выбрать наиболее оптимальные. Здесь же, на всякий случай, следует напомнить о такой очевидной вещи, что при сканировании страницы следует прислонять к стеклу сканера поплотнее (чтобы

изображение на сканах получалось как можно более резким и чётким), держа их при этом прямо и неподвижно. Простота получения скана хорошего качества будет зависеть от типа оптической системы в используемом аппарате – для оцифровки книг лучше подходят ПЗС-сканеры, так как они обладают большей глубиной резкости (±3 мм) по сравнению КДИ-устройствами (±0,3 мм) – как легко догадаться, первые менее требовательны к плотности прилегания сканируемой страницы.

Результат сканирования в конечном счёте записывается в виде графического файла. Поскольку форматов их существует великое множество, а программы управления сканером, как правило, предоставляют пользователю определённый выбор, то в Таблице 2 приводятся рекомендуемые. Общий принцип, по которому выбирались указанные в ней форматы – стремление по возможности сохранить полученную со сканера информацию в неизменённом виде. По этой причине ни в коем случае не советую сохранять сканы в виде файлов форматов, использующих при сохранении изображения алгоритмы сжатия с потерями (к таковым относится формат JPEG). Вообще говоря, типы графических форматов, их характеристики (в том числе и упоминавшийся выше термин «глубина цвета»), достоинства и недостатки – это тема для отдельного разговора, способного увести далеко от создания электронных книг, поэтому настоятельно рекомендую дополнительно ознакомиться хотя бы с имеющейся в Интернете информацией по этому вопросу[1].

Таблица 2. Форматы файлов для сохранения результатов сканирования

Режим сканирования	**Графические форматы (перечислены в порядке уменьшения предпочтительности)**
Цветной	TIFF (без сжатия), BMP
Оттенки серого	TIFF (без сжатия), BMP, GIF
Чёрно-белый	BMP, TIFF (можно с типом сжатия “PackBits” или “LZW”), GIF

Заканчивая повествование о получении электронных копий страниц книги, хочу упомянуть, что помимо сканирования существует ещё один способ, по моему мнению менее приемлемый, но иногда используемый – применение цифрового фотоаппарата, когда страницы книги просто фотографируются, а полученные снимки далее копируются на компьютер. Поскольку фотографии нередко сохраняются в виде файлов JPEG, имеющих к тому же разрешение 72 или 96 dpi (наиболее часто встречающиеся значения разрешения компьютерных мониторов), то качество получаемых таким способом фотокопий страниц (особенно если у фотоаппарата нет системы стабилизации изображения) не очень высоко. Это порой весьма негативно сказывается и на качестве самой электронной книги, из этих файлов созданной.

1 Страница «Компьютерная графика. Обучающий комплекс. Полезные ссылки» // MARKLV.NAROD.RU: сайт Львовского М.Б «Информатика в школе». URL: http://marklv.narod.ru/inf/cograf.html (дата обращения: 11.10.2022)

Обработка сканов

После того, как сканы получены, желательно заняться дальнейшей обработкой графических файлов для улучшения восприятия их содержимого либо человеком (при просмотре с экрана компьютера), либо специальными компьютерными программами (о них речь пойдёт в пункте «Распознавание сканов»). Поэтому характер обработки зависит, прежде всего, от того планируется ли дальнейшее распознавание текста книги или предполагается лишь сохранение её в виде djvu-файла.

Настоятельно советую хранить отдельно файлы исходных, необработанных сканов и ни в коем случае не удалять их до тех пор, пока процесс создания электронной книги не будет доведён до конца, а ещё лучше даже после этого какое-то время сканы не уничтожать. По этой причине необходимо иметь носитель (как правило, это винчестер компьютера) соответствующей ёмкости, присутствие на котором сотни-другой файлов размером порядка 20-30 МБ каждый не будет слишком обременительным и не приведёт к острой нехватке свободного места.

Для обработки сканов в принципе сгодится любой приличный графический редактор, но я настоятельно рекомендую “Adobe Photoshop”, поскольку у него есть одна очень замечательная функция, именуемая пакетной обработкой файлов, к тому же – хорошо реализованная. Она как нельзя кстати подходит для выполнения однотипных операций (групп действий) с большим количеством файлов изображений.

В Интернете можно найти информацию о том, как пользоваться этой самой пакетной обработкой[2], хотя сложного в ней ничего нет – я в своё время сумел разобраться самостоятельно. Суть этой функции Photoshop’а в следующем. После открытия программой редактируемого файла в палитре “Actions” («Действия») можно включить запись совокупности действий над изображением (ей предварительно задаётся какое-нибудь имя), а затем выполняются необходимые операции. После окончания записи эту совокупность действий можно многократно использовать в отношении файлов, расположенных в какой-нибудь папке. Для этого нужно командой меню File → Automate → Batch... (в руссифицированной версии это Файл → Автоматизация → Пакет...) открыть диалоговое окно “Batch” («Серия»), указать в нём название применяемой совокупности действий, папку с обрабатываемыми файлами и папку, в которую нужно сохранить результат. “Photoshop” автоматически обработает указанные ему файлы, после чего останется только проверить итог его работы. Общая длительность процесса определяется мощностью компьютера и количеством обрабатываемых файлов-сканов.

2 Статья Елисеевой Н. «Пакетная обработка фотоизображений в Adobe Photoshop» // PHOTOSIGHT.MOY.SU: Фото сайт для любителей и профессионалов. URL: https://photosight.moy.su/forum/20-10-1 (дата обращения: 11.10.2022)

Пакетная обработка – очень мощное и весьма удобное средство. Ниже приводятся несколько примеров операций со сканами, которые можно использовать в такой обработке изображений страниц оцифровываемой книги.

Пример 1. Обрезка нежелательных полей сканов.

Большинство моделей сканеров рассчитаны на сканирование страниц формата A4, при этом размер стекла сканера и, соответственно, область, с которой производится получение изображения, бывает несколько больше. В результате все сканы страниц книг могут иметь нежелательные поля, подобные тем, что схематично показаны на Рисунке 2. Поскольку поля эти редко выглядят привлекательно, увеличивая попутно размер файла скана, то их лучше удалять, используя в пакетной обработке после выделения нужного участка изображения команду меню Image → Crop (Изображение → Кадрировать…) или инструмент «Кадрирование»  на соответствующей панели. Применение данного типа обработки актуально, если предполагается создание книги в формате DJVU.

Рисунок 2. Схематичное изображение примера нежелательных правого и нижнего тёмных полей на скане (слева) и результат их удаления (справа)

Пример 2. Удаление «грязи» на полях и в центре разворота.

На сканах, полученных в монохромном режиме работы сканера, часто остаётся «грязь» в виде чёрных областей и линий по краям скана и в его середине – в месте, где проходит книжный корешок (в случае, если скан является изображением книжного разворота), что иллюстрируется на Рисунке 3. Если предполагается создание книги в формате DJVU, то имеет смысл от подобных «загрязнений» избавляться, поскольку в этом случае страница получаемой книги выглядит более опрятно, к тому же при выводе такой страницы на печать будет достигаться ещё и некоторая экономия тонера.

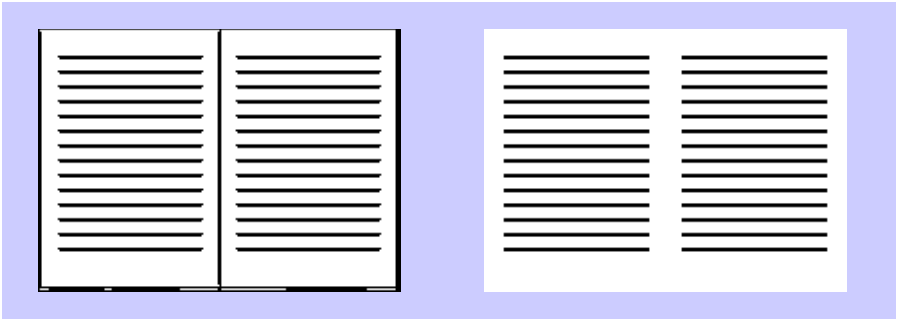
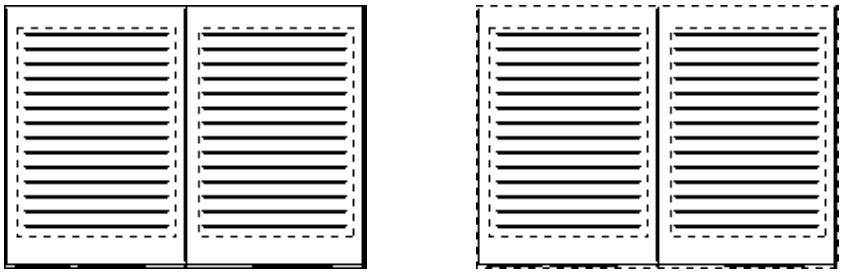


Рисунок 3. Схематичное изображение «грязи» на чёрно-белом скане (слева) и результат её удаления (справа)

Применительно к изображённому схематично на Рисунке 3 скану разворота книги выполнять очистку довольно удобно следующим образом: инструментом «Выделение» нужно выделить две прямоугольные области  (при этом на панели «Настройки» для данного инструмента должна быть активна опция “Add to selection” («Добавить к выделению»)  с текстовыми блоками страниц книги, как это показано на Рисунке 4 (а), а затем выполнить команды меню Select → Inverse (Выделение → Инвертировать выделение) (Рисунок 4, (б)) и Edit → Clear (Редактирование → Очистить).



(а) (б)

Рисунок 4. Очистка скана: выделение прямоугольных блоков текста (а) и инверсия области выделения (б)

Пример 3. Преобразование скана, полученного в оттенках серого в чёрно-белое (монохромное) изображение.

Если книга не содержит на своих страницах цветных и полутоновых изображений, то рациональнее проводить сканирование в чёрно-белом (монохромном) режиме, поскольку это благотворно скажется на размере готового файла djvu-книги. Но иногда может оказаться, что сканы страниц такой книги были получены в режиме оттенков серого или даже в цветном. В этом случае пакетная обработка может помочь в преобразовании сканов в чёрно-белые изображения. Для этого обычно бывает достаточно включить в неё команду меню Image → Adjustments → Threshold (Изображение → Регулировки → Порог...). Она «отсекает» блеклое изображение фона, заменяя его белым цветом, и делает буквы текста чёрными, чётко очерченными. Предварительно иногда бывает необходимо несколько увеличить контрастность командой Image → Adjustments → Brightness/Contrast (Изображение → Регулировки → Яркость/Контраст…). Примерный получающийся при этом результат приведён на Рисунке 5.

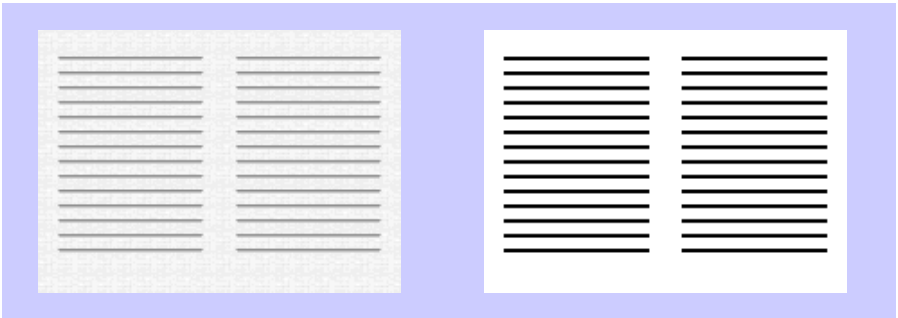


Рисунок 5. Преобразование скана полученного в режиме «оттенки серого» (слева) в чёрно-белое изображение (справа)

Здесь же мне хотелось бы добавить описание следующего приёма, многими практикуемого. Иногда книгу, не содержащую цветных и полутоновых изображений, специально сканируют в режиме оттенков серого с разрешением 300 dpi, а перед действием Threshold (Порог...) предварительно увеличивают их разрешение до 600 dpi. Делается это в диалоговом окне, вызываемом командой Image → Image Size... (Изображение → Размер изображения...), где нужно поменять разрешение (“resolution”) с 300 на 600 pixels/inch (пиксели/дюйм – эту единицу измерения можно считать тождественной dpi). Получающийся при этом обработанный рисунок неотличим от изображения страницы, изначально отсканированной в чёрно-белом режиме с разрешением 600 dpi. Таким образом можно получить книгу, как бы всю оцифрованную с указанным разрешением, но при этом сэкономить массу времени, так как нередко сканеры в режиме grayscale/300 dpi работают заметно шустрее, чем в monochrome/600 dpi.

Пример 4. Цветокоррекция сканов.

Зачастую после сканирования книги полученные сканы нуждаются в определённой цветокоррекции, одинаковой для каждого из них. Для этих целей в Photoshop’е есть целый арсенал команд из меню Image → Adjustments (Изображение → Регулировки), часть из которых уже упоминалась в предыдущих примерах. Дать конкретные рекомендации в этом случае затруднительно, поскольку применение тех или иных инструментов (команд) зависит от ситуации, от того насколько сильно и в какую сторону фактическое изображение отсканированной страницы отличается от ожидаемого результата. Так, если предполагается дальнейшее распознавание текста в сканах (для книг форматов PDF и CHM), то, пожалуй, следует обратить внимание на контрастность изображения, а если на сканах имеются цветные иллюстрации, то, возможно, понадобятся кое-какие манипуляции с балансом цветов.

Пример 5. Разделение скана с разворотом книги на отдельные страницы.

Встречающиеся в Интернете книги формата DJVU можно условно разделить на две группы: те, в которых каждой странице djvu-файла соответствует скан одной страницы книги, и на те, в которых одна djvu-страница представляет собой книжный разворот (две страницы книги). Некоторые программы для просмотра djvu-файлов имеют опцию показа сразу двух страниц, чем имитируется демонстрация книжного разворота, однако не у всех приложений эта функция всегда работает корректно. В связи с этим нет единого мнения по поводу того, как лучше сохранять книгу – постранично или разворотами – это определяется вкусом и пристрастиями создателя электронной версии какого-либо издания.

Если всё же при создании книги в формате DJVU принято решение в пользу постраничного варианта, а сканы представляют изображения книжных разворотов, то пакетной обработкой легко выполнить такое разрезание каждого скана на две картинки. Действовать тут

можно аналогично тому, как это описано выше в Примере 1, только здесь нужно записать две разные совокупности действий: одна будет кадрировать и сохранять в отдельный файл левую часть скана с изображением чётной страницы книги (см. Рисунок 2), а другая – вырезать правую его часть с изображением нечётной страницы. Нужно обязательно проконтролировать, чтобы имена получившихся файлов соответствовали порядку страниц книги. Рекомендую после выполнения одной совокупности действий, например, когда из исходных сканов получены чётные страницы, дать файлам с ними имена, содержащие чётные числа.

Для этой цели может пригодиться “Total Commander” или во многом аналогичный ему “Free Commander”. У данных приложений есть инструмент “Multi Rename” по переименованию сразу целой группы выбранных файлов – он находится в меню File (Файл). При его вызове открывается диалоговое окно, в котором можно добавить к началу имени каждого файла в папке число (или вообще присвоить имя, только из этого числа и состоящее). Для этого предварительно в шаблоне новых имён файлов указывается элемент “Counter” («Счётчик»), у которого нужно будет задать следующие дополнительные параметры:

- начальное значение счётчика (“Start at:” – «Начать с:»), в зависимости от ситуации – 0, 1 или 2;
- шаг (“Step by:”) – при нумеровании чётных или нечётных страниц-файлов установить значение 2;
- число знаков (цифр) (“Digits:”) – зависит от общего числа страниц в отсканированной книге. Так. если их количество исчисляется сотнями, то лучше поставить значение 3.

После переименования можно запускать на выполнение другую совокупность действий (вырезание из сканов нечётных страниц книги), по завершении которой понадобится выполнить ещё одно аналогичное переименование. В результате файлы с отдельными страницами книги можно будет спокойно поместить в одну папку, и все они там расположатся в правильном порядке.

Иногда возникают ситуации, в которых пакетная обработка не в состоянии помочь. Например, на одной из страниц книги может оказаться чья-нибудь рукописная пометка типа «Вася Пупкин – дурак» или рисунок в стиле «точка, точка, запятая – вышла рожица смешная». Если после сканирования книги с такими дополнительными «декорациями» нет желания оставить их красоваться в djvu-файле, то следует выполнить ретуширование скана, дабы удалить из него следы «народного творчества». В Photoshop’е подобная процедура проводится с использованием инструментов «Размытие»  , «Клонирование штампа»  и некоторых других. Занятие это требует терпения и аккуратности, поэтому вполне способно отнять довольно много времени, особенно если нежелательная «левая» надпись сделана прямо поверх

текста книги (наличие такой надписи приведёт к его некорректному распознаванию в этом месте).

Распознавание сканов

По завершении сканирования и обработки сканы представляют собой изображения, при просмотре которых на экране компьютера любому человеку по силам не только определить, что перед ним – именно текст, но и прочитать его. Однако для компьютера эта картинка с буквами – всего лишь картинка, и текст в ней, так легко идентифицируемый человеком, для нашего электронного друга «мёртв», машина его не «видит» и, следовательно, не может с ним ничего делать, например, осуществлять полнотекстовый поиск. Однако эта проблема на сегодняшний день вполне разрешима благодаря существованию приложений для распознавания текста, которые ещё называются OCR-программами (OCR – optical character recognition – «оптическое распознавание текста»). Распознавание – это особый вид обработки изображения в графическом файле, заключающийся в преобразовании элементов изображения в последовательность текстовых символов. Полученный в результате текст уже является «компьютерно-живым»: его можно копировать, редактировать или преобразовывать каким-нибудь ещё способом при помощи любых текстовых редакторов.

Существующий сегодня ассортимент OCR-программ довольно обширен. Наиболее известной здесь является ABBYY “Fine Reader”, однако существуют и другие программные продукты, например, бесплатный “CuneiForm”. Стоит добавить, что в состав программного обеспечения, входящего в комплект поставки сканеров, тоже часто входят программы-распознавалки от фирмы-разработчика сканера. Функция распознавания текста есть и у программы “OmniPage”, упоминавшейся в пункте «Сканирование».

Так уж сложилось, что все OCR-программы «знают» английский язык, в связи с чем при распознавании текстов на нём проблем не возникает. С языком русским дело обстоит сложнее, поэтому при выборе программы для распознавания именно русского текста первоочередным критерием должно служить наличие возможности настроить её на восприятие нашего «великого и могучего». Перечисленные абзацем выше приложения (“Fine Reader”, “CuneiForm”, “OmniPage”) текст на русском понимают достаточно хорошо, правильно определяя начертание кириллических символов.

Для распознавания программе нужно указать файл или диапазон файлов. Как правило, OCR-приложения умеют работать с довольно разнообразными графическими форматами (TIFF, GIF, BMP, JPEG). У некоторых приложений есть такая особенность: они считывают метаданные из файла скана, определяют разрешение изображения, и если оно оказывается меньше 300 dpi, то выдаётся предупреждающее сообщение о возможных ошибках распознавания текста на такой картинке. Практика же показывает, что для успешного распознавания главенствующую роль

играет не разрешение, а то насколько чётко на скане прорисованы буквы и их характерные элементы.

Распознанный текст можно сохранить в виде отдельного файла. Многие OCR-программы здесь также предлагают целый набор возможных форматов: TXT, HTML, RTF и даже DOC. Я бы рекомендовал первый из перечисленных вариантов, ввиду его наибольшей универсальности. Кроме этого, при сохранении распознанного текста в других форматах к нему добавляется ещё и выбираемое программой форматирование, которое не всегда оказывается уместным и в дальнейшем может даже мешать.

Формат DJVU тоже позволяет сохранять в файле книги распознанный текст. Более подробно об этом будет рассказано в пункте «Создание выходного файла».

Корректура

После того, как в процессе распознавания получен файл (файлы) с текстом книги, необходимо выполнить его корректуру – исправить ошибки в нём, поскольку они возникают почти всегда, даже если проводилось распознавание сканов отличного качества. И хотя количество ошибок может быть совсем небольшим, но всё-таки ещё ни одному OCR-приложению не удалось превзойти в этом плане зоркий глаз человека.

Строго говоря, «корректура» – понятие типографское и в действительности оно имеет значительно более широкий смысл, нежели простая вычитка распознанного текста. Однако здесь под корректурой, прежде всего, будет пониматься процесс выявления и ликвидации ошибок работы OCR-программы. В этом случае рассматриваемый термин совпадает по значению с использующимся иногда англоязычным понятием “spell-checking” (переводится как «проверка орфографии»).

Как правило, для исправления ошибок распознанный текст переносят в какой-нибудь текстовый редактор и проводят правку в его среде. К сожалению, часто ограничиваются только теми исправлениями, которые предлагаются самим редактором – в современных программах этого типа имеются встроенные возможности по проверке орфографии и грамматики. Так, MS “Word” и LO “Writer” подчёркивают красной волнистой линией «подозрительные» с их точки зрения слова, которые после этого сразу бросаются в глаза. Однако не стоит забывать, что далеко не от всех ошибок можно избавиться подобным образом. После программной проверки следует провести ещё одну, хотя и затратную по времени, но значительно более действенную. Для этого нужно внимательно прочитать весь распознанный текст, используя в качестве эталона либо саму книгу, либо открытые каким-нибудь просмотрщиком сканы её страниц.

Занятие это не из лёгких, поскольку компьютерный текст сильнее утомляет глаза, нежели печатный, по причине чего при чтении довольно быстро снижается внимание и многие малозаметные ошибки в тексте (например, «бараика» вместо «баранка») могут так и остаться

неисправленными. Здесь можно рекомендовать следующий подход: проверяемый текст нужно размеренно читать вслух по слогам, стараясь чётко их проговаривать. Это позволяет сконцентрироваться именно на побуквенном написании слов, а не на их общем виде (зрительном образе). Следует добавить, что при таком способе контроля текст воспринимается не только глазами, но ещё и на слух, а это также облегчает выявление ошибок.

Не стоит забывать, что и в исходных текстах книг могут присутствовать опечатки – для современных изданий это, к превеликому сожалению, стало особенно актуальным. Если в процессе проверки распознанного текста в оригинале книги действительно будет выявлена опечатка, то, по моему мнению, её нужно исправить. В дальнейшем, в готовом (выходном) файле книги, необходимо будет обязательно оставить информацию обо всех внесённых исправлениях, подобно тому, как в бумажных книгах иногда делается небольшая вклейка с перечнем невыправленных опечаток.

При создании книги в формате DJVU также вполне возможна кое-какая корректорская правка, но в этом случае вам понадобится графический редактор вместо текстового, поскольку править придётся непосредственно сканы. Далее в формируемую электронную книгу нужно будет непременно добавить дополнительную страницу со сведениями о том, какие именно изменения и в каком месте книги были внесены, туда можно также добавить и фрагмент скана исходной страницы с неисправленной опечаткой.

Заканчивая описание этого этапа создания электронной книги, хочу обратить внимание на следующее. Поскольку правила правописания, которые мы все изучаем в школе, имеют обыкновение подзабываться со временем, то нелишним будет при проведении вычитки текста дополнительно вооружиться справочными источниками – онлайновыми[3] или печатными (например, таким: *Розенталь Д.Э. Справочник по правописанию и литературной правке: Для работников печати. – 5-е изд., испр. – М.: Книга, 1989. – 320 с.*).

Подготовка иллюстративного материала

Как известно, djvu-книги представляют собой особым образом сжатые изображения страниц издания-оригинала. По этой причине если в исходной книге имелись рисунки, то необходимости выделения их из текста не возникает. Совсем другое дело, если создаётся pdf- или chm-книга с распознанным текстом, но в ней должны присутствовать изображения, иллюстрирующие её содержимое. В этом случае необходимо выделить из сканов страниц оригинала все рисунки, схемы, диаграммы и дополнительно их обработать. Здесь снова понадобится графический редактор – “Photoshop” или “GIMP”.

3 Раздел «Словари» на портале Грамота.ру // GRAMOTA.RU: Справочно-информационный портал Грамота.ру. URL: http://gramota.ru/slovari/ (дата обращения: 11.10.2022)

На мой взгляд, наиболее оптимальным при подготовке иллюстраций является следование принципам, которые используются в веб-дизайне при подготовке файлов изображений, размещаемых в Интернете. В подавляющем большинстве случаев это файлы форматов JPEG, GIF или PNG. Сказанное вовсе не означает, что можно просто вырезать рисунки, диаграммы, графики из сканов страниц книги и сохранять их в одном из указанных форматов. К этому желательно подходить рационально.

Во-первых, поскольку страница отсканирована скорее всего с разрешением 300 dpi (а может быть и больше), то в пиксельном выражении все рисунки будут довольно большими. В связи с этим необходимо уменьшить их ширину и высоту, приблизив размеры рисунка при отображении на компьютерном мониторе к фактическим, то есть к тем, какие он имел в оригинале на книжной странице. Следует оговориться, что уменьшать рисунки нужно осмотрительно – иногда они содержат много мелких деталей, которые при изменении размера могут просто потеряться, в результате информативность изображения снизится.

Во-вторых, для выделенного из скана рисунка может понадобиться цветокоррекция (если она не была проведена ранее – см. пункт «Обработка сканов») для улучшения его вида.

В-третьих, перед сохранением готового файла с картинкой для дополнительного уменьшения его размера нужно оптимизировать изображение. Например, в Photoshop’е для этого служит инструмент File → Save for Web... (Файл → Сохранить для Web...). Подробный рассказ о тонкостях работы с ним выходит за пределы тематики настоящего пособия, поэтому рекомендую изучить этот вопрос самостоятельно, используя имеющуюся в Интернете информацию.

Если книжная страница отсканирована в монохромном режиме, то имеющиеся на ней и предварительно «вырезанные» из скана рисунки можно «пережимать» следующим образом:

- Перевести режим изображения (Image → Mode – «Изображение → Режим») из монохромного (Bitmap – «Битовый») в оттенки серого (Grayscale – «Черно-белый») или RGB;
- Уменьшить разрешение до 72 dpi;
- Выполнить команду Image → Adjustments → Posterize... (Изображение → Регулировки → Постеризовать...). В появившемся диалоговом окне установить значение Levels (Уровни) в интервале от 4 до 8 (в зависимости от того, насколько сильно «огрубляется» изображение);
- Перевести режим изображения в индексированные цвета (Image → Mode → Indexed Color... – «Изображение → Режим → Индексированные цвета»), после чего сохранить файл в формате GIF (File → Save As... – «Файл → Сохранить как...»), выбрав при этом опцию “Normal”.

При действии описанным способом получается файл довольно небольшого размера. На Рисунке 6 показан пример исходного рисунка и результат такой его обработки.

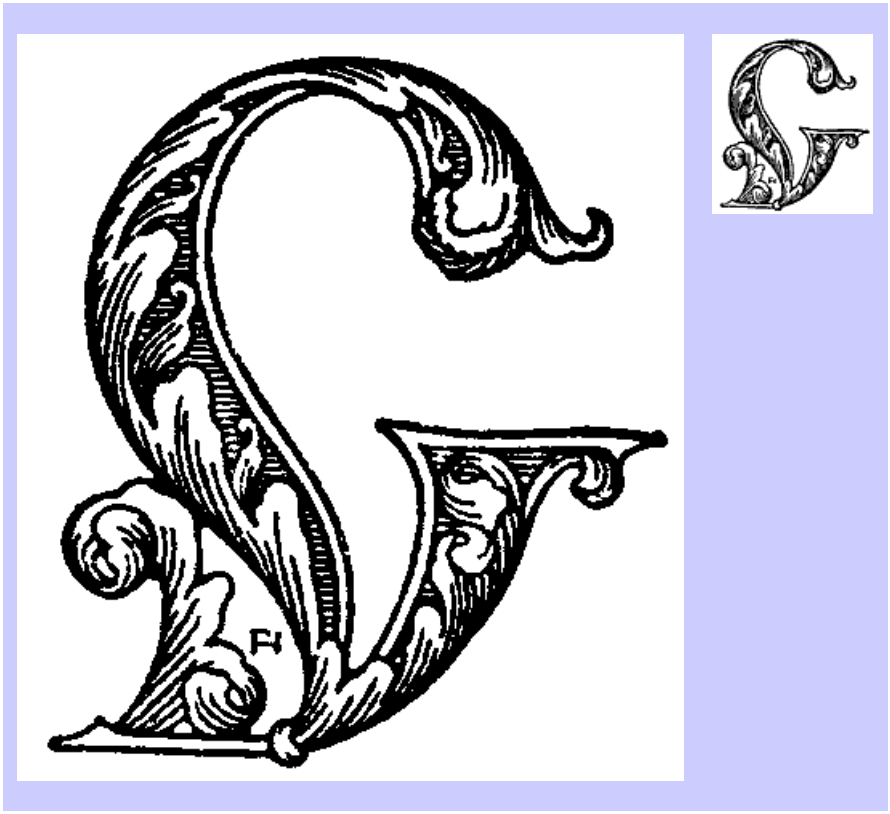


Рисунок 6. Исходное изображение на скане страницы (слева) и оно же после пережатия (справа). В качестве примера использована буквица на с. 6 издания: *Живое слово. Книга для изучения родного языка. Часть II / Сост. Острогорский А.Я. 10-е изд. – Петроград: Типография Тренке и Фюсно, 1916. – 399 с.*

Во многих книгах встречается представление данных в виде таблиц. В электронных же версиях таких изданий часто эти таблицы вставлены в текст книги в виде рисунков (фрагментов сканов), обработанных аналогично тому, как это описано выше. Некоторые OCR-программы (например, “CuneiForm”) умеют определять, что на скане имеется таблица и позволяют сохранить распознанную информацию именно в виде таблицы (электронной). И хотя она, скорее всего, будет нуждаться в дополнительной проверке её данных и форматировании при помощи пригодной для этого программы (MS “Word”, “Excel” и т. п.), после распознавания табличная информация станет, как и остальной текст, «компьютерно-живой» – её данные смогут быть легко переданы (скопированы) во многие другие программы. Если таблиц в книге немного и они не очень объёмные, то их можно набрать вручную, что также обеспечит их «компьютерную живость» в готовой книге. На мой взгляд, такой подход более предпочтителен, нежели просто вставка сканированного изображения таблицы в оцифрованную книгу.

Ещё одна разновидность иллюстративного материала, встречающаяся в книгах – это различные схемы (например, блок-схемы алгоритмов в учебниках по программированию). Если планируется создание электронной книги в формате PDF, а схем, подобно вышеописанному случаю с таблицами, не очень много и они довольно простые, то имеет смысл не оставлять

каждую из них растровой картинкой, а перерисовать, представив в виде векторного рисунка. Создать его можно при помощи любого редактора векторной графики, которым вы владеете (LO “Draw”, например). Вполне могут сгодиться и штатные средства работы с векторной графикой, имеющиеся в MS “PowerPoint”, чтобы нарисованную вручную схему сохранить затем в формате WMF.

Вёрстка

Книги на бумаге человечество печатает уже не один век, и со времён Иоганна Гуттенберга (в Европе) и Ивана Фёдорова (в России) в этой области деятельности выработан целый комплекс предписаний и правил, направленных на создание эстетичного вида страниц и лёгкость восприятия печатного слова. Эти принципы могут быть в определённой степени распространены и на слово «компьютерное».

Когда текст электронной книги и её иллюстративный материал подготовлены, можно приступать к следующей стадии – вёрстке. Как и корректура, это тоже типографский термин. Вёрстку обычно определяют как процесс составления страниц из текста и иллюстраций в соответствии с выбранным заранее способом оформления материала (дизайном). Из такого определения следует одно из главных требований, которому должна удовлетворять вёрстка – единообразие страниц.

Сейчас при подготовке печатных изданий широко применяется вёрстка при помощи компьютера. Для этого существуют профессиональное и, как следствие, малопонятное непосвящённому человеку программное обеспечение. Если вы владеете навыками работы в среде подобных программ, то при создании электронных книг используйте их, новичкам же лучше предложить варианты попроще – именно о них и будет далее рассказано, но при этом совсем не возбраняется дополнительно просветиться по части типографской вёрстки и дизайна[4, 5].

Если планируется создание книги в формате PDF, то её вполне прилично можно подготовить и при помощи хорошего текстового редактора. Для этого подойдут неоднократно уже упоминавшиеся MS “Word” или LO “Writer”. Эти программы сами проводят распределение текста по строкам (автоматическая вёрстка строк) и по страницам (автоматическая вёрстка страниц), Кроме этого, они способны следить, чтобы в формируемом документе не было так называемых «висячих строк» и других, нежелательных с типографской точки зрения, вещей. О том, как скомпонованный в виде файла DOC/DOCX или, соответственно, ODT (собственный

4 Статья «Верстка книг – начнем с азов» // DPK-PRESS.RU: Издательство «ДПК Пресс». URL: https://dpk-press.ru/verstka-knigi-nachnem-s-azov/ (дата обращения: 11.10.2022)

5 Лебедев А. «Ководство» // WWW.ARTLEBEDEV.RU: Студия Артемия Лебедева. URL: https://www.artlebedev.ru/kovodstvo/sections/ (дата обращения: 11.10.2022)

формат текстовых документов Writer'а) материал превратить в pdf-книгу, будет рассказано в пункте «Создание выходного файла».

Файл формата CHM является по сути заархивированным набором веб-страниц (html-файлов), а это означает, что перед созданием chm-книги нужно сначала эти веб-страницы подготовить (способы преобразования html-файлов в CHM будут указаны в пункте «Создание выходного файла»). Для этого также существуют специальные программы – html-редакторы. В принципе, в такой роли может выступить практически любой текстовый редактор, хоть стандартный windows'овский «Блокнот» ("Notepad") – теоретически с его помощью можно создать веб-страницу любой сложности, но в этом случае без знания языка HTML (HyperText Markup Language – «язык гипертекстовой разметки») никак не обойтись. Собственно, при работе и с «нормальным» html-редактором знание этого компьютерного языка очень желательно – сейчас изучение его основ даже включено в школьный курс по информатике: *Угринович Н.Д. Информатика и информационные технологии. Учебник для 10-11 классов. М.: БИНОМ. Лаборатория знаний, 2003. – 512 с. (Глава 13 «Основы языка гипертекстовой разметки документов», с. 467)*.

В Word'е веб-страница может быть создана командой Файл → Сохранить как… , где в качестве формата выходного файла надо лишь выбрать «Web-страница» или «Web-страница с фильтром».

Сама вёрстка несколько различна при подготовке файлов форматов PDF и CHM (HTML), однако есть некоторые общие моменты, которые следует иметь в виду как одном, так и в другом случае.

1. Для достижения единообразия в оформлении книги лучше всего использовать для её содержимого стилевую разметку, поэтому в начале выполнения вёрстки должна быть разработана система стилей. В текстовых редакторах есть специальные средства для работы с ними, при подготовке веб-страниц удобно пользоваться CSS (Cascading Style Sheets – «каскадные таблицы стилей»).

Применительно к области электронных документов понятие «стиль» означает заранее оговорённую и поименованную совокупность свойств (атрибутов) текста – начертание символов (гарнитура), их размер (кегль), величина междустрочного интервала (интерлиньяж), выравнивание абзаца (выключка) и т. д. Стиль позволяет сразу задать части электронной книги всё необходимое форматирование, а не щёлкать многократно мышью, указывая, что, например, какой-то конкретный фрагмент текста должен быть набран шрифтом "Arial" размером 10 пунктов, междустрочный интервал у него должен быть полуторным, абзацный отступ – 1 см, а абзацы иметь выравнивание по ширине.

Нужно заранее определиться, сколько различных стилей потребуется для оформления книги, какие наборы свойств будут иметь основной её текст, названия глав, подзаголовков,

подписи к иллюстрациям (если таковые имеются) и прочее, стараясь при этом свести количество требуемых стилей к минимуму. Следует помнить, что стиль – это не только способ быстрого присваивания нужного форматирования, но и средство эффективного управления оформлением и задания структуры содержимого подготавливаемой книги. Применение стилевой разметки в среде текстового редактора замечательно тем, что при изменении какого-нибудь атрибута (свойства) в стиле это автоматически сказывается на внешнем виде материала, которому этот стиль присвоен. Таким образом, если материал книги размечен, то очень легко, например, поменять при необходимости размер шрифта (кегль) во всём тексте или сделать написание всех заголовков курсивным (если оно было прямым) вне зависимости от того сколько их – десять или все сто.

Задание структуры книги в текстовых редакторах реализуется за счёт особых стилей заголовков разного уровня. Сходную роль при разработке веб-страниц играют теги заголовков языка HTML – `<H1>...</H1>`, `<H2>...</H2>` и т. д. Благодаря присвоенным стилям заголовков компьютер начинает «видеть» в подготавливаемом файле (например, DOC), какая глава книги из скольких подзаголовков состоит и какой объём текста в конкретной главе (пункте, параграфе). Именно это обстоятельство позволяет реализовать автоматическое составление книжного оглавления, избавляя пользователя от необходимости вручную контролировать правильность простановки номеров страниц в нём.

2. В отличие от бумажных книг, таблицы – это не только средство для сжатого и лаконичного представления порой достаточно большого количества данных, дополняющих и иллюстрирующих основной текст книги, но и эффективный, а зачастую – незаменимый инструмент для вёрстки, в частности для размещения на странице в определённом порядке разнородного материала. При подготовке веб-страниц для сайтов такой подход применяется регулярно. При оформлении документов в форматах DOC или ODT использование таблиц также может оказаться весьма удобным. Взгляните на Рисунок 7 – на нём приведён пример, когда иллюстрации и текст на странице размещены довольно затейливым образом, но достигнуто это простым распределением материала по ячейкам таблицы 3×3.

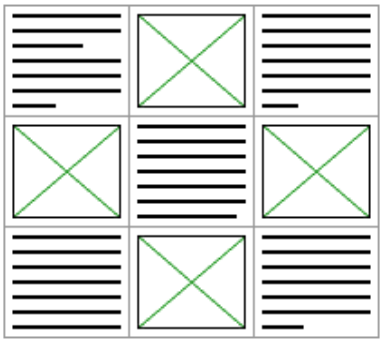


Рисунок 7. Пример использования таблицы при совместной вёрстке на странице рисунков и текста. Для большей наглядности у ячеек линии границ, которые в таких случаях должны быть «невидимыми», специально показаны серым цветом

3. Важное значение имеют правила набора текста. Для печатных изданий, как отмечалось выше, такие требования давным-давно разработаны и используются. В случае электронных публикаций хотя и нет правил, утверждённых в виде ГОСТов, стандартов ИСО и т. п., но всё же многое из того, что относится к печатной продукции, неплохо транслируется и на продукцию электронную, правда не без некоторой специфики. При подготовке электронных книг желательно придерживаться традиционных правил набора, поэтому всем интересующимся рекомендую к ознакомлению информацию как из соответствующих онлайн-ресурсов[6], так и из специализированного печатного издания: *Гиленсон П.Г. Справочник художественного и технического редакторов. М.: Книга, 1988. – 528 с.* – думаю, что информации в указанных источниках будет вполне достаточно. Перечислять правила я не буду, поскольку хочу сделать акцент на другом: зачастую они только требуют, что нужно делать, не проливая света на то, как именно это делать при помощи программных средств. Вот несколько примеров.

Известно, что должны отделяться друг от друга пробелом: фамилия и инициалы (Пупкин В.), значение какой-либо величины и единицы измерения (60 км/ч), символ номера и число (№ 13). Если для этого использовать обычный пробел, то бывают случаи, что фамилия, значение величины или символ номера оказываются в конце одной строки, а, соответственно, инициалы, единицы измерения или число – в начале следующей. Чтобы этого избежать (то есть, чтобы одно не «отрывалось» от другого), нужно ставить символ «неразрывный пробел». В Writer’е и Word’е он вставляется сочетанием клавиш Ctrl + Shift + Пробел, в веб-страницах это достигается добавлением в html-код в нужном месте комбинации символов (так называемой мнемоники) ` ` .

Есть сложные слова, пишущиеся через дефис: Ростов-на-Дону, ёлки-палки, Олгой-Хорхой. Подобно вышеприведённому примеру, браузеры и текстовые редакторы могут часть слова оставить на одной строке, а часть – перенести на другую. Ничего особенного здесь нет, но бывают ситуации, в которых текст, написанный через дефисы, не должен подвергаться такому переносу, например, это относится к телефонным номерам. В Word’е в подобных случаях следует использовать символ «неразрывный дефис». Поставить его можно командой Вставка → Символ… . В языке HTML для подобных случаев есть специальный тег `<NOBR>`: `<NOBR>Тел.: 987-65-43</NOBR>`.

Позволяют расширять возможности по оформлению и другие особые непечатаемые знаки – так в MS “Word” и LO “Writer” есть группа «разрывных» символов, например, «Разрыв страницы» – принудительный перенос последующего текста в начало новой страницы (можно

6 Статья Шестакова А.П. «Правила компьютерного набора текста» // COMP-SCIENCE.NAROD.RU: Сайт «Учителям информатики и математики и их любознательным ученикам (дидактические материалы по информатике и математике)». URL: http://comp-science.narod.ru/pr_nab.htm (дата обращения: 11.10.2022)

использовать комбинацию клавиш Ctrl + Enter), или «Разрыв строки» – принудительный перенос слова в абзаце в начало следующей строки (сочетание Shift + Enter).

В текстах печатных книг помимо обычных символов время от времени встречаются такие знаки, которых нет на клавиатуре: греческая буква «пи» π, символ «промилле» ‰, параграф § и т. п. Изредка, но можно встретить, что в электронных версиях изданий спецсимволы представлены в виде «суррогатов». Так, знак копирайта "©" имитируют сочетанием символов "(C)", а стрелку влево "←" – при помощи комбинации "<–". Для книг в формате PDF и CHM (HTML) это сейчас совершенно неприемлемо, поскольку текстовые редакторы имеют инструменты, позволяющие достаточно легко вставлять спецсимволы, а в языке HTML используются мнемоники (для упоминавшихся выше знаков они следующие: π – `&pi;`, ‰ – `&permil;`, § – `&sect;`, © – `&copy;`, ← – `&larr;`).

Пользователи Интернета прекрасно знают, что современные сайты чрезвычайно разнообразны по своему внешнему виду и встречаются очень красиво и удобно оформленные сетевые ресурсы. Довольно оригинально бывают оформлены и некоторые chm-книги, поэтому, в дополнение к вышесказанному, следует отметить, что и при подготовке веб-страниц для файла формата CHM также следует придерживаться традиций, успевших уже устояться в среде такой ещё относительно молодой отрасли как веб-дизайн – информацию на эту тему легко отыскать в сети.

Создание выходного файла

Ну вот мы и подошли к рассмотрению финальной стадии изготовления электронных книг, и все предыдущие этапы можно рассматривать лишь как подготовительные, что, впрочем, нисколько не умаляет их важности. Начнём с DJVU.

Существует целый ряд приложений, позволяющих создавать и редактировать djvu-файлы. Я рекомендую довольно простую в использовании программу “DjvuSolo” версии 3.1. Файлы формата DJVU бывают разных модификаций (версий). DjvuSolo создаёт файлы достаточно ранней версии 21, которые открываются любыми приложениями для просмотра DJVU, чем обеспечивается максимальная совместимость. Считаю нужным предупредить, что поскольку указанное приложение уже довольно старое, с его работоспособностью в среде новых версий ОС Windows могут возникать проблемы.

Процедура создания djvu-книги довольно проста: нужно открыть в программе файл скана, добавить к нему остальные изображения со страницами книги и перекодировать всё это командой File → Encode As DjVu. “DjvuSolo” умеет открывать графические файлы в форматах BMP, JPEG, GIF, TIFF и некоторые другие, при этом необходимо помнить, что сканы зачастую

ввиду их большого размера открываются достаточно долго – возможно придётся запастись терпением, ожидая, пока приложение их «прожуёт».

В процессе программа дополнительно просит указать пару настроек.

Во-первых, она спрашивает, в каком виде сохранить результат: одним единым файлом, вместе (“Bundled”), или же раздельно, как группу djvu-файлов (“Indirect”), каждый из которых содержит изображение одной страницы (скана).

Во-вторых, нужно также указать метод, в соответствии с которым будет производиться перекодирование изображения на сканах в формат DJVU. В “DjvuSolo” этих методов четыре:

- “Scan” – наиболее подходит для изображений страниц, полученных в режимах работы сканера «Цветной» или «Оттенки серого». При кодировании производится анализ изображения, в результате которого малоконтрастные участки с плавными переходами цветов сохраняются в слой “background”, а высококонтрастная часть картинки – текст, графики, таблицы – сохраняется в слои “mask” (само изображение) и “foreground” (данные о цвете для слоя трафарета-маски).
- “Clean” – в чём-то аналогичен “Scan”, но предназначен главным образом для перекодирования изображений, имеющих «компьютерное» происхождение (например, скриншотов рабочих окон программ), поскольку в этом режиме отделение текста от фона выполняется более аккуратно. Тем не менее, этот метод для отсканированных книжных страниц сколько-нибудь заметного преимущества (размер выходного файла, качество картинки) по сравнению с методом “Scan”, как правило, не даёт.
- “Photo” – всё изображение сохраняется целиком в слой “background” и в результате сильно проигрывает в размере выходного файла по сравнению с методом “Scan”, хотя по качеству ему особо не уступает. Для сканов книжных страниц использовать в принципе можно, но очень осмотрительно.
- “Bitonal” – обрабатываемое изображение предварительно переводится в чёрно-белый режим, а затем сохраняется в слой “mask”. Этот метод лучше всего использовать для сохранения сканов, полученных в монохромном режиме работы сканера.

Часто бывает, что сканируемая книга содержит на некоторых страницах цветные иллюстрации, на некоторых – чёрно-белые (полутоновые серые), а также страницы только с текстом. Такие книги рациональнее сканировать с применением всех трёх режимов сканирования. При создании из сканов djvu-файла лучше сначала цветные и полутоновые серые страницы сохранить в одном файле с использованием метода “Scan”, а монохромные – в другом

файле, использовав метод “Bitonal”. Далее содержимое этих файлов можно при помощи той же “DjvuSolo” объединить в одну djvu-книгу.

Алгоритм, в соответствии с которым происходит перекодирование изображения скана страницы в DJVU, имеет одну неприятную особенность: он сам может генерировать «опечатки», что как раз и обусловлено отбрасыванием по мнению алгоритма «лишней» информации при сжатии скана. Наиболее часто встречающаяся ошибка – это замена буквы «и» буквой «н» и наоборот. Подобные искажения в процессе сжатия исходной картинки могут возникать у текста, набранного мелким шрифтом или на сканах плохого качества (когда изображение нерезкое, несколько размытое). Именно по этим причинам в пункте «Сканирование» специально было обращено внимание на правильный выбор разрешения и на плотное прилегание страницы к стеклу сканера.

Djvu-файлы могут содержать в себе слой распознанного текста, что порой бывает весьма удобно, поскольку отпадает необходимость экспорта страниц и обработки с помощью OCR-программы, но, как нетрудно догадаться, наличие дополнительной информации приводит к некоторому увеличению размера файла. Функция распознавания русскоязычного текста есть, в частности, в шестой версии djvu-редактора “Document Express Editor” – в ней это достигается командой меню Tools → OCR → OCR Entire Document (Сервис → OCR → OCR документа).

Формат DJVU дозволяет создавать в электронной книге гиперссылки в виде активных областей страницы, имеющих прямоугольную, овальную или полигональную форму. Ссылки могут направлять пользователя либо на какой-нибудь интернет-ресурс, либо на другую страницу многостраничного djvu-файла. Последнее позволяет сделать в такой книге интерактивное оглавление, поэтому не стоит лениться создавать его. За основу можно взять страницу (страницы) со сканом настоящего оглавления книги и «оживить» его при помощи гиперссылок – пользоваться такой электронной книгой удобнее, поскольку в ней читателю проще добраться до интересующего его материала.

Для создания файлов в формате PDF есть несколько способов. Можно использовать LO “Writer” – для этого достаточно открыть этой программой файл DOC или ODT с подготовленной книгой и воспользоваться командой Файл → Экспорт в → Экспорт в PDF… . Кроме этого, есть ряд программ, так называемых «виртуальных принтеров», например “doPDF”. Пользоваться подобными приложениями также несложно: преобразуемый в pdf-книгу файл нужно просто отправить на печать через такой «принтер» – программа перехватит выводимую информацию и запишет в виде файла – «напечатанная» виртуальным принтером страница выглядит также, как она выглядела бы, будучи напечатанной на бумаге принтером физическим (реальным).

Текст из pdf-файла можно поместить (скопировать) в буфер обмена, однако с кириллическими знаками могут возникнуть неприятности – иногда русский текст копируется, но вставляется в виде совершенно нечитаемой каши из непонятных символов-закорючек («кракозябр»). С pdf-файлами, созданными LO “Writer” эта проблема решается так: перед выделением копируемого текста индикатор раскладки клавиатуры должен быть переключен на «Русский» (RU). Кстати, при копировании распознанного русскоязычного текста из файла формата DJVU описанная сложность тоже может проявиться – решается она аналогично.

Дополнительно хотелось бы рассмотреть такой вопрос. Ещё в пункте «Форматы книг» упоминалось, что встречаются pdf-книги, представляющие собой набор отсканированных страниц, а в пункте «Общие положения» говорилось, что такие электронные версии изданий уступают по размерам аналогичным файлам DJVU. Вот один из способов, с помощью которого можно преобразовать книгу в формате PDF в djvu-файл – зачастую при этом удаётся достичь весьма существенного уменьшения размера файла с книгой. Для этого вам пригодится “PDFill” или “FinePrint” – это тоже виртуальные принтеры. Рассматриваемые приложения умеют отправленную на «печать» информацию сохранить в виде набора графических файлов популярных форматов (прежде всего – TIFF и BMP). Таким образом, из pdf-файла можно получить как бы набор «сканов» со страницами книги. Далее эти «сканы», в зависимости от ситуации, можно с использованием пакетной обработки в Photoshop’е подрезать (удалить лишние поля), преобразовать, допустим, из оттенков серого в монохромные и т. п. По завершении этих процедур останется лишь перекодировать обработанные «сканы» в DJVU.

Для создания файлов CHM есть разные программы. Прежде всего, это “HTML Help Workshop” от Microsoft, создательницы этого формата. Я рекомендую также программу “Htm2chm”, позволяющую легко конвертировать (компилировать) набор подготовленных предварительно веб-страниц с сопутствующими им файлами в один. Одно из неоспоримых достоинств данного приложения – исключительно простой и интуитивно понятный интерфейс, который позволяет очень быстро освоить программу.

Не могу не упомянуть приложение “ChmBookCreator”, которое можно бесплатно скачать с сайта «Маленькая паутинка» (“smallweb.ru”). Данное приложение позволяет создавать приятно оформленные chm-книги непосредственно из файлов RTF, DOC и некоторых других. При этом “ChmBookCreator” обладает рядом настроек, позволяющих пользователю дополнительно управлять оформлением создаваемой электронной книги.