Инструкция по созданию электронных книг из бумажных оригиналов

версия 1.1. от 16.12.2017

Замечания общего характера

Процесс сканирования книг и последующей обработки сканов является длительным и оттого зачастую нудным, но все-таки творческим занятием. От того, как аккуратно ведется сканирование и обработка книг, в значительной степени зависит качество результата. Если вы уже собрались потратить несколько часов своей жизни на то, чтобы отсканировать какую-то книгу, то не обесценивайте собственную работу – сделайте всё качественно. Помните, что у книги, которая уже есть в электронном виде в интернете, шансов быть отсканированной повторно почти нет, т.к. книг на свете много, а силы человеческие не безграничны. Поэтому если вы выложите в интернет ваш отвратительный скан, то вы тем самым не только снизите удовольствие от книги читателям, но и резко сократите шансы на то, что когда-нибудь эта книга будет оцифрована с приличным качеством. Уважайте себя и других.

Предлагаемая инструкция позволит вам создавать электронные книги в форматах Djvu или Pdf с четкими буквами, ровными строчками, текстовым слоем и активным оглавлением, а также картинками высокого качества. Примеры готовых электронных книг, созданных по этой технологии, можно найти на сайте http://vas-s-al.livejournal.com/ по тегу «книга».

Данная инструкция составлена на основе личного опыта и не претендует на всеобщность. Я вполне признаю, что могут существовать неизвестные мне программы и технические приемы, которые позволяют делать всё то же самое проще или с более высоким качеством. Поэтому я буду признателен вам за любые советы о том, как можно улучшить и упростить процесс.

И пожалуйста, не забудьте выложить электронную книгу в интернет.

Часть 1. Сканирование

Оборудование и программы

Для оцифровки книги необходим сканер или фотоаппарат. Оцифровку книги фотоаппаратом мы разберем в разделе «сканирование в полевых условиях», а пока будем считать, что книга находится у вас на руках, и вы можете использовать стационарный сканер, подключенный к компьютеру. Я использую специализированный книжный сканер **Plustek OpticBook 3800** и абсолютно им доволен. Это единственный известный мне книжный сканер, который стоит относительно приемлемые деньги (около 20 тысяч рублей по состоянию на конец 2017 года). Отличается он от обычного сканера только тем, что область сканирования (стекло) у него смещена к краю, что позволяет класть на него книгу «уголком», не разворачивая ее полностью, и сканируя таким образом не разворот, а каждую страницу. За счет этого во-первых удается избегать оптических искажений, которые неизбежно возникают у корешка книги, если сканировать ее разворотом, во-вторых, сама книга остается более целой, т.к. многие издания от раскрывания их на 180 градусов могут просто развалиться, а в-третьих становится возможным сканировать книги формата А3, которые в развернутом виде в обычный сканер просто не влезут.

Кроме того, у этой модели удобные кнопки управления, каждая из которых отвечает за свой режим сканирования (цветной, в градациях серого и черно-белый). Обычно я настраиваю сканер так, чтобы в цвете он сканировал с разрешением 600 DPI, а в градациях серого – с разрешением 300 DPI. Это позволяет не перенастраивать сканер в процессе и ускоряет работу.

Рисунок 1. Сканер Plustek OpticBook 3800 и принцип его действия.

Конечно, неплохо бы иметь полностью автоматизированный книжный сканер, который умеет самостоятельно переворачивать страницы, но такие аппараты обычно стоят как хороший автомобиль, поэтому их могут позволить себе только какие-нибудь серьезные организации. Что характерно, самые отвратительные по качеству электронные книги обычно являются продуктом деятельности сотрудников государственных библиотек, что в очередной раз доказывает, что никакие деньги и никакой официальный статус не могут заменить наплевательского отношения к работе.

Если у вас обычный бытовой сканер, то это значит лишь то, что вам придётся в процессе сканирования сильнее прижимать к стеклу вашу книгу, и книги большого формата в него не влезут.

Если вы созрели для покупки сканера, но не созрели покупать агрегат за 20 тыс. рублей (хотя на Авито можно найти дешевле), то при покупке поинтересуйтесь типом светочувствительного элемента (бывают двух видов: CCD и CIS). У CIS неплотно прижатые страницы размываются, поэтому для книг нужен только CCD (рисунок 2).

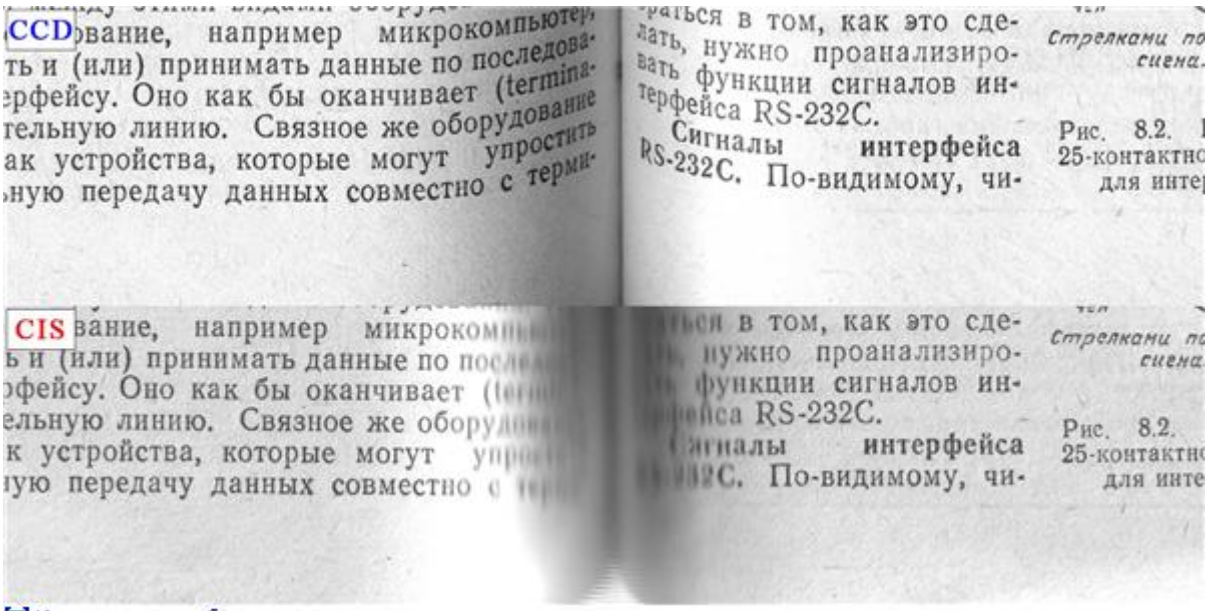


Рисунок 2. Сравнение сканов неплотно прижатой книги сканерами с разными типами сканирующих элементов (картинка с сайта http://www.djvu-soft.narod.ru).

Ваш сканер должен быть способен сканировать с разрешением не ниже 600 DPI, а жесткий диск должен иметь несколько гигабайт свободного места. В принципе, это все требования к «железу».

Кроме сканера и компьютера вам для сканирования изображений также пригодится лист черной бумаги. Черную бумагу можно купить в магазинах для художников или найти в детском наборе для аппликаций. На худой конец, можно просто напечатать на любом принтере «чёрный квадрат». Лист бумаги мы будем подкладывать под страницу с изображением с обратной стороны, чтобы сквозь светлые области изображения не просвечивали буквы на обороте.

Еще одним «оборудованием», которое всегда должно быть под рукой, является тряпочка из микрофибры, которую можно купить в любом хозяйственном магазине. С книг на стекло сканера летит пыль, а если книга старая, то и кусочки бумаги, и для того, чтобы всего этого не было на скане, сканер нужно периодически протирать тряпочкой от пыли. Важно знать, что пыль не только может оставлять следы на скане, но и нарушить нормальную работу самого сканера.

Иногда при сканировании в цветном режиме на скане начинают появляться тонкие вертикальные цветные полоски (рисунок 3). Дело в том, что на сканирующей головке обычно приклеена полоска белой бумаги, и сканер периодически калибрует по ней сам себя, определяя, что такое «белый», и на этой основе воспринимая уже все остальные цвета. Если на эту полоску попадет пыль (которая, само собой, не совсем белая), то в том месте, где осела пылинка, сканер будет принимать за белый какой-то другой цвет, и, соответственно, цветопередача нарушится. Лечится эта проблема аккуратным (аккуратным!) разбором сканера и обметанием этой белой полоски мягкой чистой кисточкой.

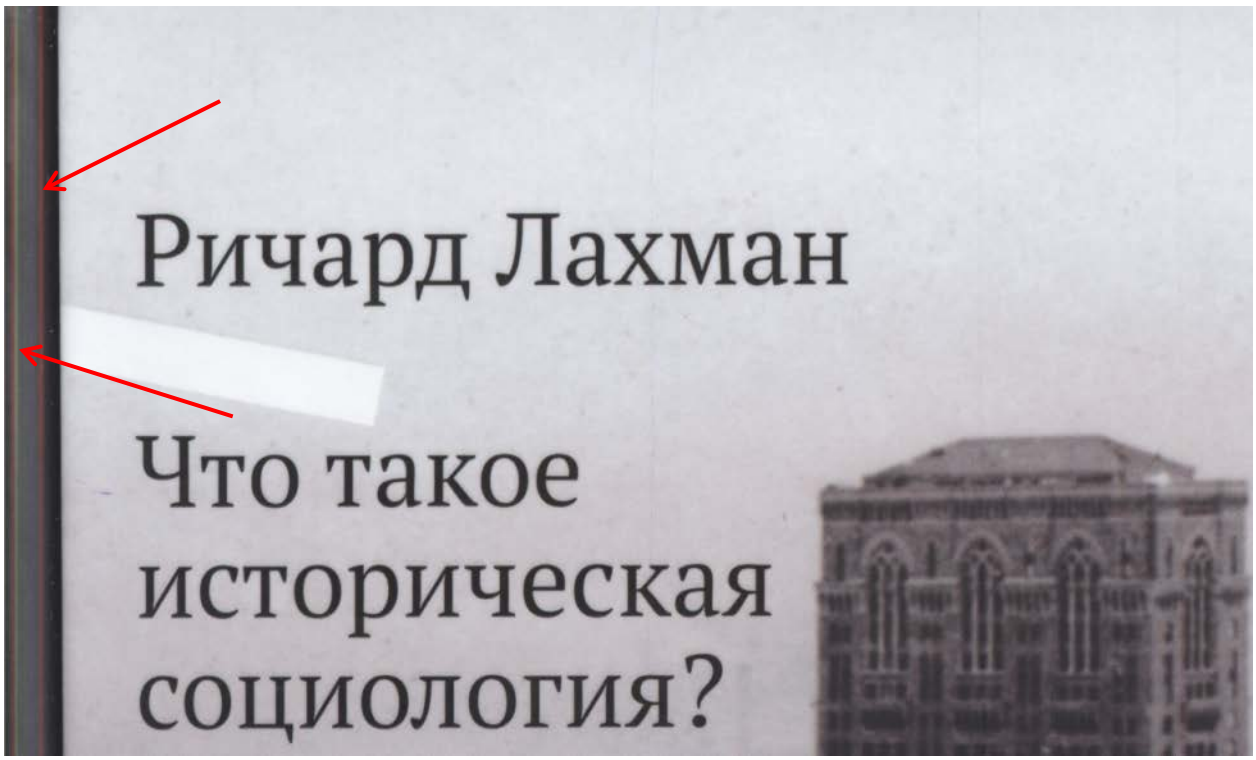


Рисунок 3. Цветные полосы на скане из-за попавшей в сканер пыли.

Я пользуюсь для сканирования встроенной утилитой сканера, в которой можно задавать формат сохранения сканов, разрешение сканирования, путь для сохранения результатов и автоматический поворот. Если вы будете использовать обычный бытовой сканер, то, возможно, его программное обеспечение (ПО) не будет столь удобным. Вы можете пользоваться любой программой для сканирования, важно лишь помнить некоторые требования, которым она должна отвечать.

Программа должна иметь возможность отключения всех автоматических «улучшателей» изображения. Зачастую скан на экране монитора выглядит более контрастным, более насыщенным, чем страница книги. Это происходит из-за того, что производители сканеров пытаются подменять своими программами Adobe Photoshop и другие нормальные редакторы изображений, как правило, делая это неудачно и топорно. К сожалению, в последней версии драйвера для Plustek OpticBook 3800 производитель тоже не избежал этого греха, поэтому, когда я обновил драйвер, сканы стали получаться какими-то слишком контрастными. К счастью, у меня сохранилась старая версия драйвера без «улучшателей». Если у вас тоже Plustek OpticBook 3800, вы можете использовать этот https://yadi.sk/d/xSOP4YuSSaW7q драйвер для Windows 8 и выше.

Помимо возможности отключения всех «улучшений», выбранная вами программа для сканирования должна иметь возможность сохранять сканы в формате BMP или Tiff. При этом нужно быть внимательным – зачастую программа сохраняет файл в формате Tiff со сжатием Jpeg. Дело в том, что формат Tiff поддерживает сохранение изображения со сжатием, которое может осуществляться различными алгоритмами. Сжатие без потери качества дает использование алгоритма LZV. Однако, к примеру, моя утилита сканера сохраняет Tiff со сжатием Jpeg, причем эта функция является неотключаемой. Узнать это можно в любой программе просмотра изображений, найдя там кнопочку «сведения о файле». Сохранение в Jpeg – это сохранение с потерей качества изображения. Поэтому этот формат лучше не использовать. Нам требуется возможность сохранять изображения без потери качества. Поэтому если ваша программа также сохраняет файлы в формат Tiff, который на проверку оказывается замаскированным Jpeg, то вам, как и мне, придется сканировать всё в формат BMP, а потом конвертировать сканы в формат Tiff с правильным алгоритмом сжатия. Книжный магазин «Фаланстер» в своей инструкции по

сканированию книг рекомендует использовать для самого процесса сканирования программу AcdSee Pro, но мне хватает встроенной утилиты сканера. Инструкция «Фаланстера» доступна здесь: http://falanster.livejournal.com/199543.html

Я пользуюсь для переконвертации сканов бесплатной программой **XnView**. Ее можно скачать на сайте разработчика: http://www.xnview.com/en/xnview/

Если книга толстая, и ваших сил, когда вы будете прижимать листы к стеклу сканера, всё равно не хватит, чтобы избежать оптических искажений у корешка (строчки на скане начинают как бы «загибаться»), то вам придется на стадии обработки сканов использовать программу **Book Restorer** или встроенную функцию программы **Scan Tailor**, которая умеет распрямлять их обратно.

Процесс сканирования

Сканирование ста страниц книги обычного формата с разрешением 300 DPI занимает у меня примерно полчаса (если не отвлекаться). Таким образом, на книгу в 300 страниц придётся потратить часа полтора. Это время можно использовать для того, чтобы послушать какие-нибудь лекции в интернете или аудиокнигу.

Сканировать страницы с текстом нужно с разрешением не ниже 300 DPI, страницы с графиками, диаграммами, рисунками (особенно рисунками!) – не ниже 600 DPI. Текст можно сканировать в режиме «градации серого», рисунки – в цветном режиме. Сканирование в режиме «черно-белый» использовать нельзя!

Почему необходимо сканировать изображения с разрешением не меньше 600 DPI? Потому что изображения в книгах, как правило, состоят из множества маленьких точек, и если разрешения сканера не хватает, чтобы «увидеть» каждую из них, то при обработке картинки компьютер немного «сходит с ума», и на изображениях появляются характерные артефакты (рисунок 4).

Рисунок 4. Артефакты, появившиеся при сканировании растрового изображения с недостаточным разрешением.

Для удаления растра существуют специальные плагины для фотошопа. В простейшем случае изображение просто размывается с глубиной размытия, подобранной так, чтобы точек уже не было видно, но мелкие детали при этом ещё бы не терялись. Упомянутые плагины подбирают глубину размытия автоматически. Подробнее об обработке изображений мы поговорим во второй части инструкции, посвященной обработке сканов, а пока нужно просто запомнить, что **все страницы с изображениями необходимо сканировать с разрешением не ниже 600 DPI.**

Кроме того, для сканирования изображений стоит использовать лист черной бумаги. Подкладывайте его с оборотной стороны страницы, и тогда на изображениях не будут просвечивать напечатанные на обороте буквы.

Размеры области сканирования вы должны задать такие, чтобы с одной стороны страница или разворот книги сканировались целиком, но с другой стороны, чтобы никакие посторонние предметы (корешок книги, ваши пальцы, крышка сканера и т.п.) в область сканирования не попадали. Потом, когда мы станем обрабатывать сканы программами, которые распознают область текста, любые посторонние детали могут стать им помехой, и то, что программа обычно делает быстро и автоматически, нам придётся делать медленно и вручную.

Формат для сохранения сканов нужно установить bmp или tiff, но используйте формат tiff не раньше, чем вы убедитесь, что ваша утилита сканера не сжимает его или сжимает по алгоритму lzv. **Tiff, сжатый алгоритмом jpeg (фактически замаскированный jpeg), нам не нужен!**

Просмотреть все параметры файла, можно, нажав кнопочку свойства в программе для просмотра изображений. Я использую программу XnView, и в моем случае это выглядит так:

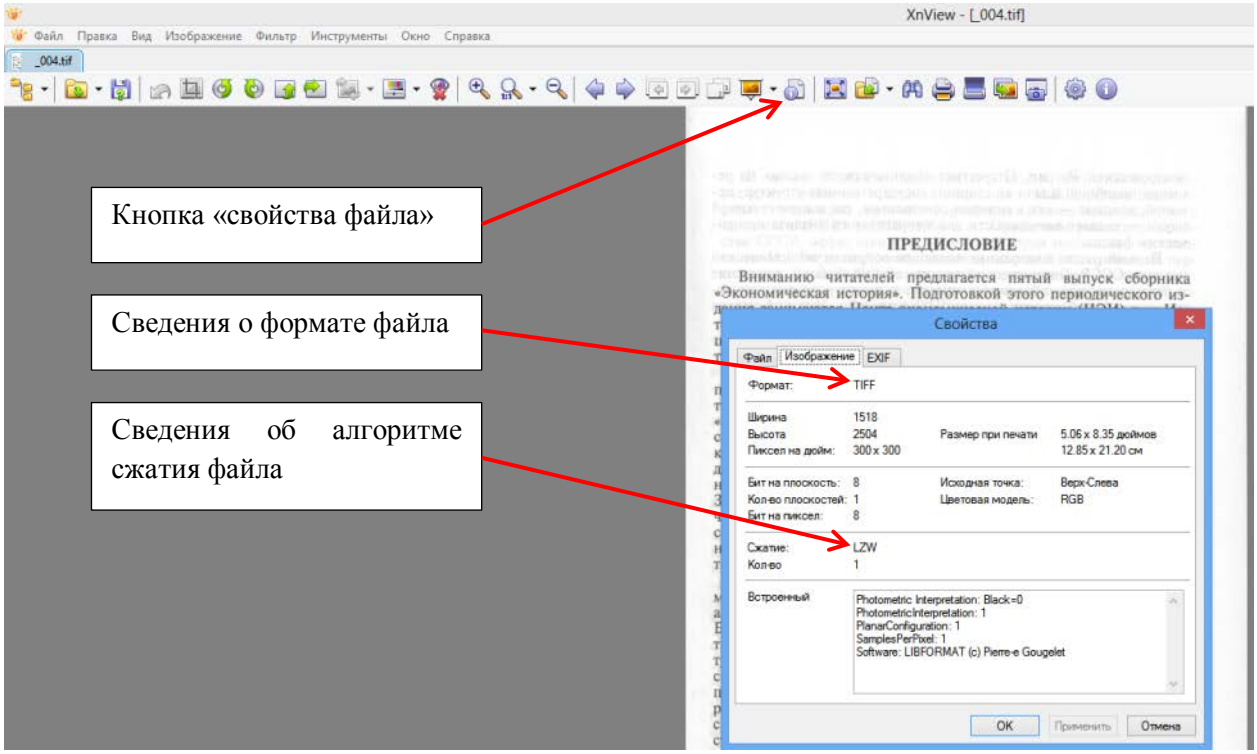


Рисунок 5. Окно «свойства файла» в программе XnView.

Для надежности лучше сканировать всё в формат BMP. Потом можно будет переконвертировать все файлы в формат tiff с нужным алгоритмом сжатия программой XnView. Переконвертация нужна не только для того, чтобы уменьшить размер пакета файлов, но и потому, что некоторые программы, которые мы будем использовать, с форматом bmp не работают.

Чтобы переконвертировать наши файлы, надо в программе XnView открыть окно «обозреватель»

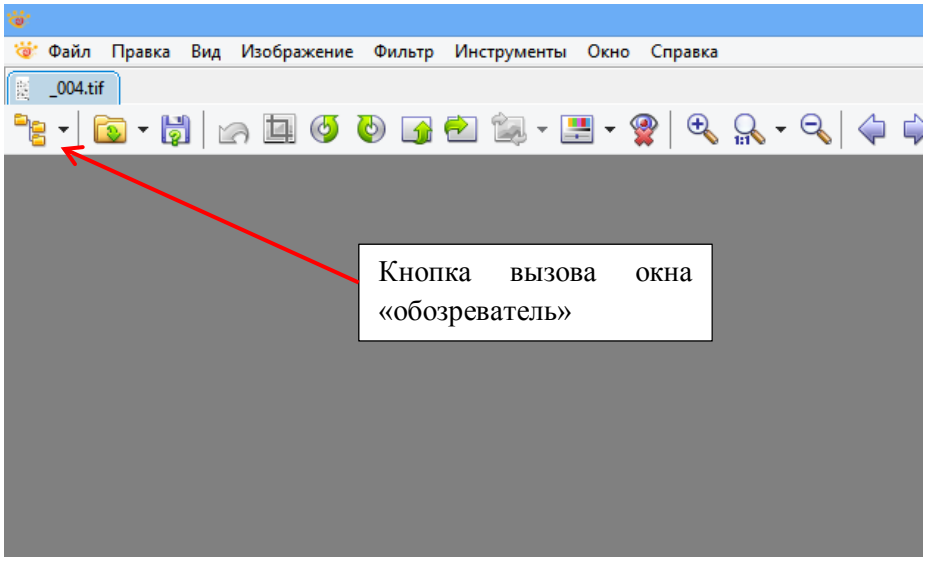


Рисунок 6. Кнопка вызова окна «обозреватель»

И затем, выбрав все нужные файлы (например, одновременным нажатием клавиш «Ctrl» и «A»), нажать кнопку «Преобразование» и в появившемся окне «Пакетная обработка» на вкладке «Основные» выбрать формат и папку сохранения, в опциях формата (кнопка «Опции» на вкладке «Основные» задать алгоритм сжатия lzv, а на вкладке «преобразования» для всех страниц, отсканированных в градациях серого, выбрать преобразование «Преобразовать в серое».

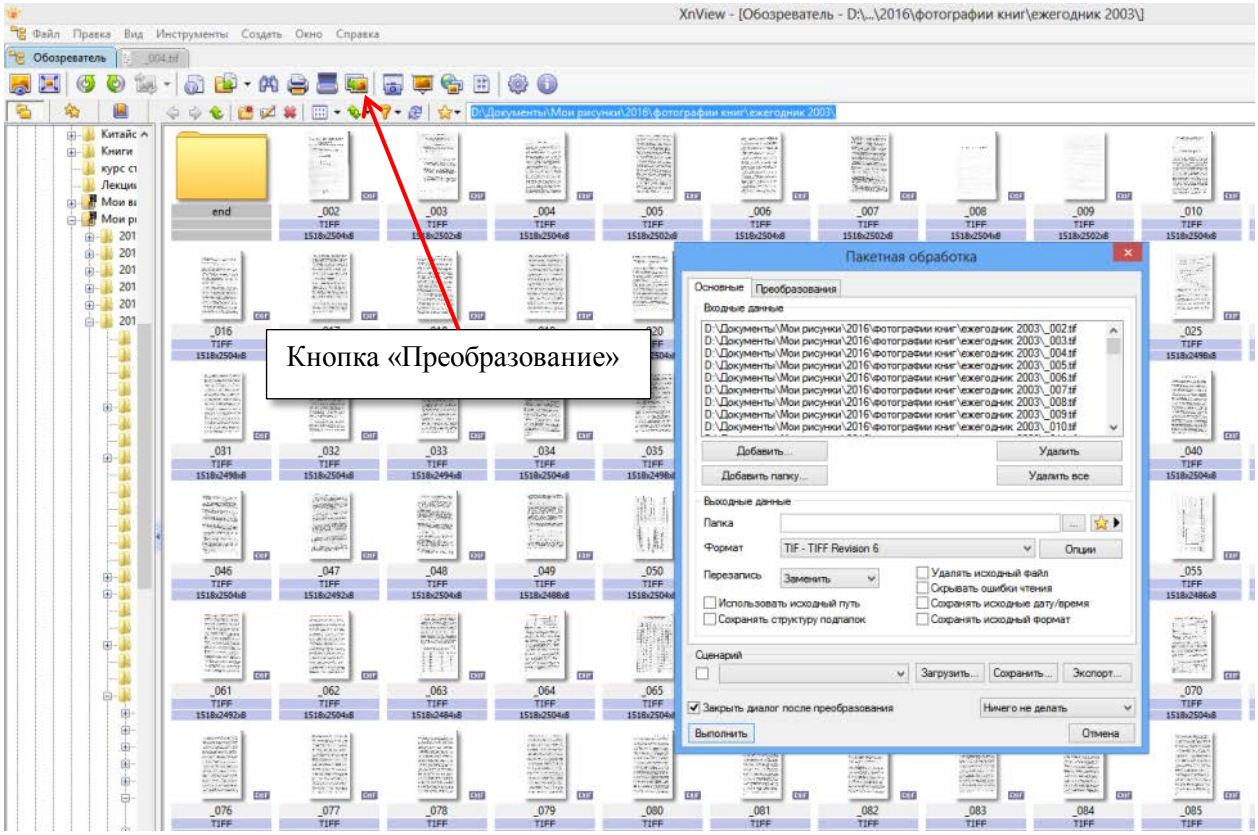


Рисунок 7. Пакетная обработка файлов в программе XnView.

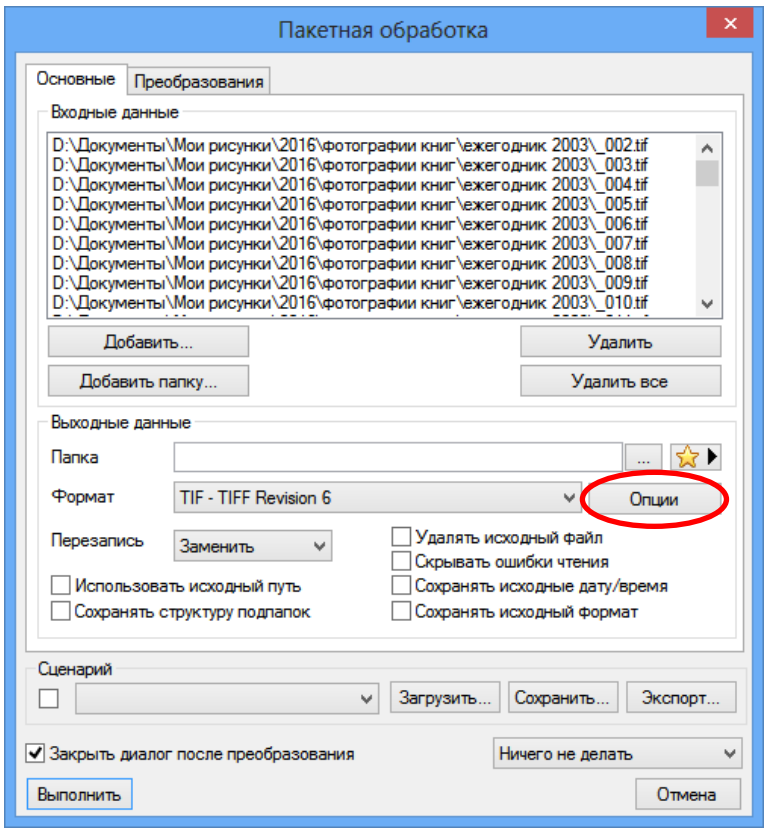


Рисунок 8. Вкладка «Основные» окна «Пакетная обработка». Рядом с форматом файла есть кнопка «Опции», нажав которую можно выбрать параметры сжатия.

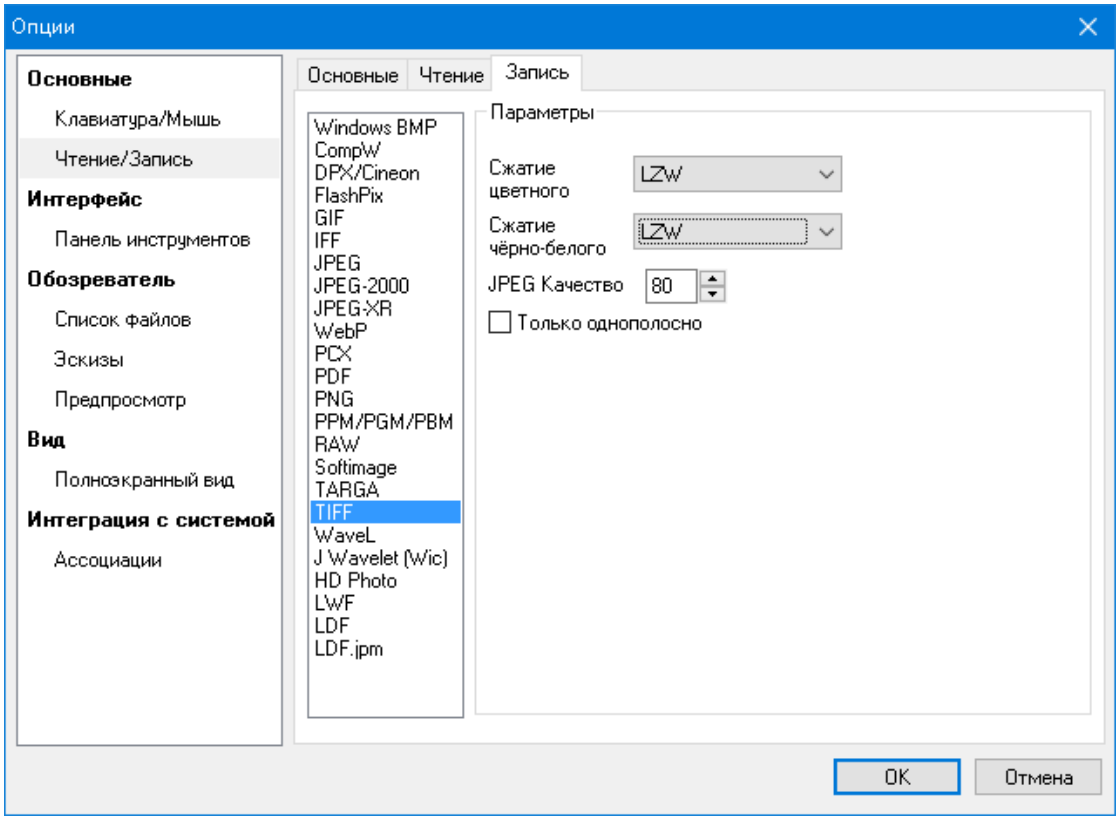


Рисунок 9. Опции сохранения (записи). Выбираем формат Tiff, параметры сжатия LZV. На параметр «Jpeg качество» внимания не обращаем, он играет роль, только если выбираем сжатие по алгоритму jpeg (чего делать не нужно).

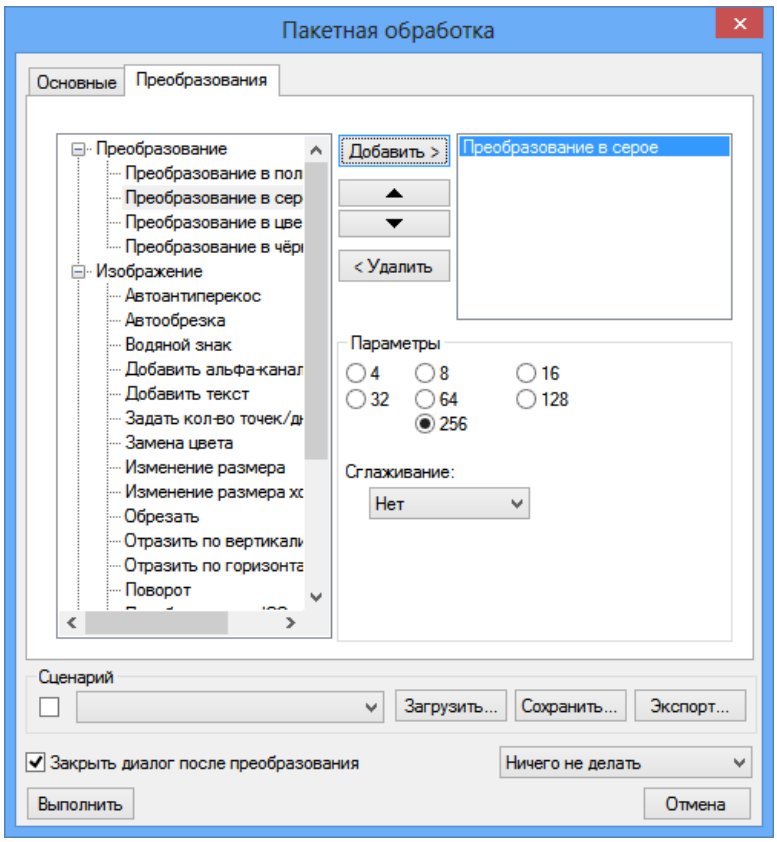


Рисунок 10. Вкладка «Преобразования» окна «Пакетная обработка».

Это удивительно, но не цветное изображение можно (и нужно!) сделать ещё более не цветным. Использование преобразования «Преобразование в серое» удалит всю информацию о цвете с отсканированных в градациях серого (grayscale) изображений и позволит в дальнейшем применить к ним очень полезные плагины в фотошопе.

Нажав кнопку «Выполнить», мы запустим пакетную обработку, и через несколько минут получим набор сканов страниц, сохраненных в правильном формате.

В большинстве случаев я сканирую книги, у которых цветными являются только страницы обложки. Но т.к. на моём сканере есть отдельные кнопки для всех режимов сканирования, то для кнопки цветного сканирования я устанавливаю разрешение в 600 dpi, а для кнопки сканирования в градациях серого – разрешение в 300 dpi, и в дальнейшем все страницы, где требуется высокое разрешение скана (а это совершенно необязательно страницы с цветными картинками) сканирую нажатием «цветной» кнопки, чтобы не тратить время на перенастройку сканера. Преобразование в серое позволяет мне удалить ненужную мне (и занимающую много места) информацию о цвете из этих страниц, сохранив их высокое разрешение.

Когда мы получили пакет файлов страниц книги в формате tiff со сжатием lzv, этап сканирования можно считать завершенным. При этом файлы должны выглядеть примерно так:

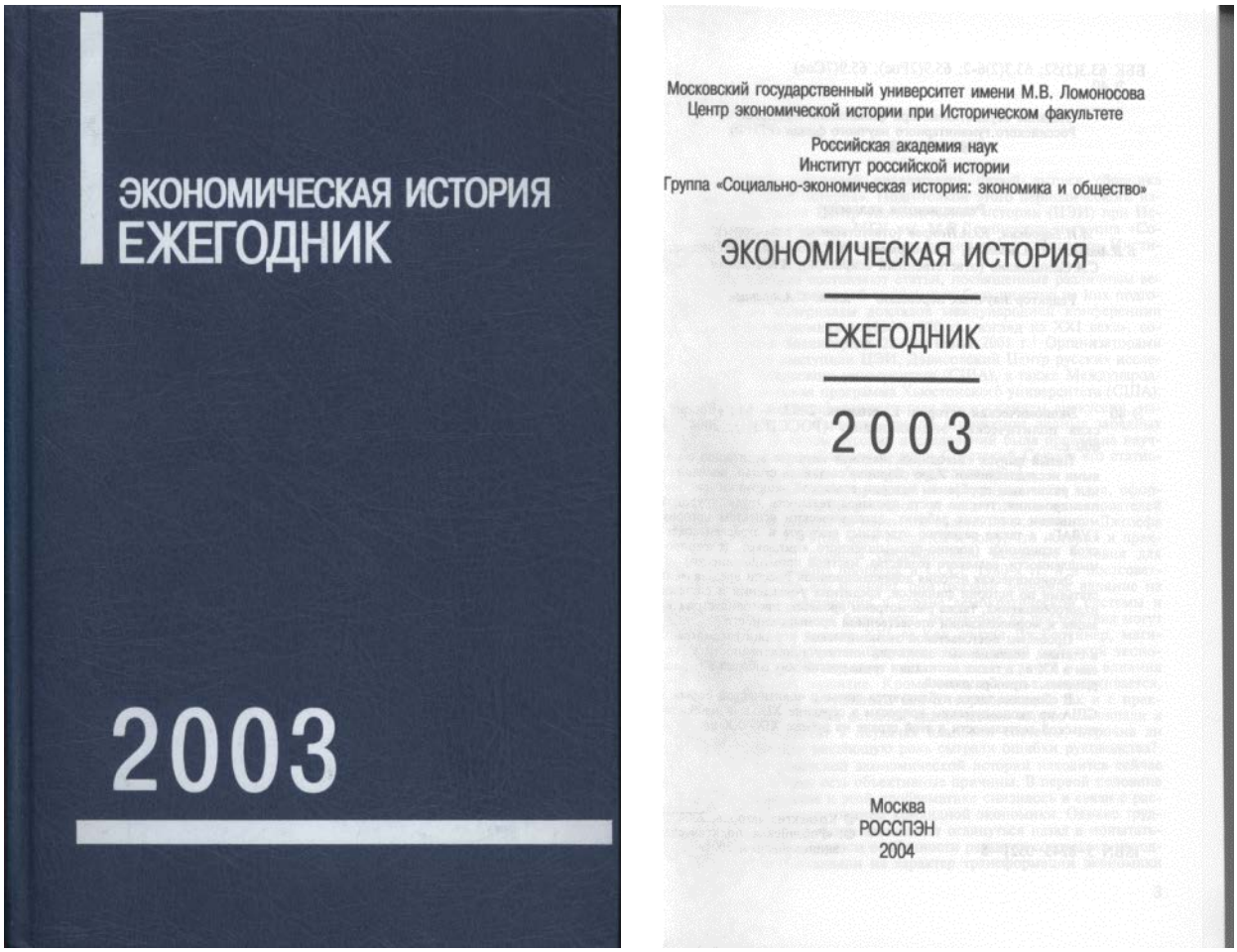


Рисунок 11. Примеры правильно отсканированных страниц – цветной и в градациях серого.

Оцифровка книг в полевых условиях

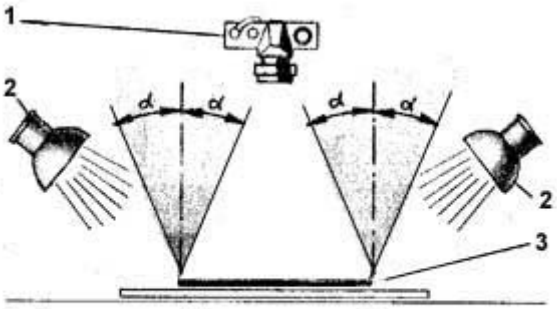
Вполне может случиться так, что нужная книга попадется вам на глаза в том месте, откуда её нельзя будет просто забрать домой и спокойно отсканировать так, как это описано выше. Я оцифровывал книги в библиотеке, музее, книжном магазине, под открытым небом на блошином рынке у букинистов, и в большинстве случаев результат получался вполне сносный.

Поэтому давайте разберем несколько простых советов, которые позволят вам фотографировать книги так, чтобы потом из ваших фотографий можно было делать нормальные читаемые PDF или DJVU. Почему фотографировать? Потому что, скорее всего, сканер из дома вы туда, где нашли книгу, всё равно не потащите. Есть, правда, ручные сканеры, которые можно носить с собой, но у меня никогда не получалось перемещать их по странице книги достаточно плавно, чтобы результат был не хуже того, который можно получить фотоаппаратом. Поэтому начнём с фотоаппарата.

1. **Фотоаппарат**. Практика показывает, что для того, чтобы книга потом нормально читалась, достаточно разрешения фотоаппарата от 6 мегапикселей. Сейчас таким разрешением обладает даже большинство телефонов, поэтому если книга нашлась совсем внезапно, то можно использовать и мобильник.
   Настройте зумом фотоаппарата приближение так, чтобы разворот книги занимал как можно большую часть кадра, но не обрезался. Фотографировать удобно стоя, положив книгу на стол перед собой. Если фотоаппарат - зеркалка, и объектив самовольно выезжает, когда вы держите фотоаппарат объективом вниз, куском скотча фиксируем нужное

фокусное расстояние. Кроме того, если ваш фотоаппарат сам умеет определять, вертикально или горизонтально вы его держите, то отключите эту функцию! Иначе замучаетесь потом разворачивать фотографии. Фотоаппарат сам на большинстве режимов подкручивает два параметра - выдержку и глубину резкости - для того, чтобы получить нормальную освещенность (экспозицию) кадра. Если вы снимаете при нехватке света, вы можете вручную изменить соотношение выдержки и глубины резкости, уменьшив глубину резкости (диафрагму) и одновременно сократив выдержку. У меня получаются достаточно четкие изображения при выдержке короче 1/30 секунды. Но чем легче фотоаппарат, тем легче он трясется и тем более короткая выдержка требуется, чтобы изображение не смазалось. Вообще зеркалка, конечно, для фотографирования предпочтительнее «мыльниц». Хотя бы потому, что у неё есть шейный ремень, и она снимает мгновенно. Если вы не знаете, что такое выдержка и диафрагма, и где в вашем фотоаппарате их настраивать, то старайтесь хотя бы не трясти фотоаппарат в момент съемки.

2. **Освещение.** Оно должно быть ярким и равномерным. Лучше всего, если это будет естественный свет солнца. Если вынести книгу на улицу нельзя, подвиньте стол к окну или положите книгу на подоконник. Если стол двигать нельзя, или всё время пасмурно (как в Москве зимой), проследите, чтобы искусственный свет был равномерным (т.е. две лампы по бокам книги, или лампа на потолке прямо над книгой). К примеру, с недавних пор в Ленинской библиотеке разрешено фотографировать. Там в читальных залах установлены столы с вмонтированными в них лампами из расчета один стол на двух читателей. Захватите целый стол, и поверните обе лампы так, чтобы они светили в центр, одна справа, другая – слева. Положите книгу в это световое пятно. Следите, чтобы на страницах не было теней.

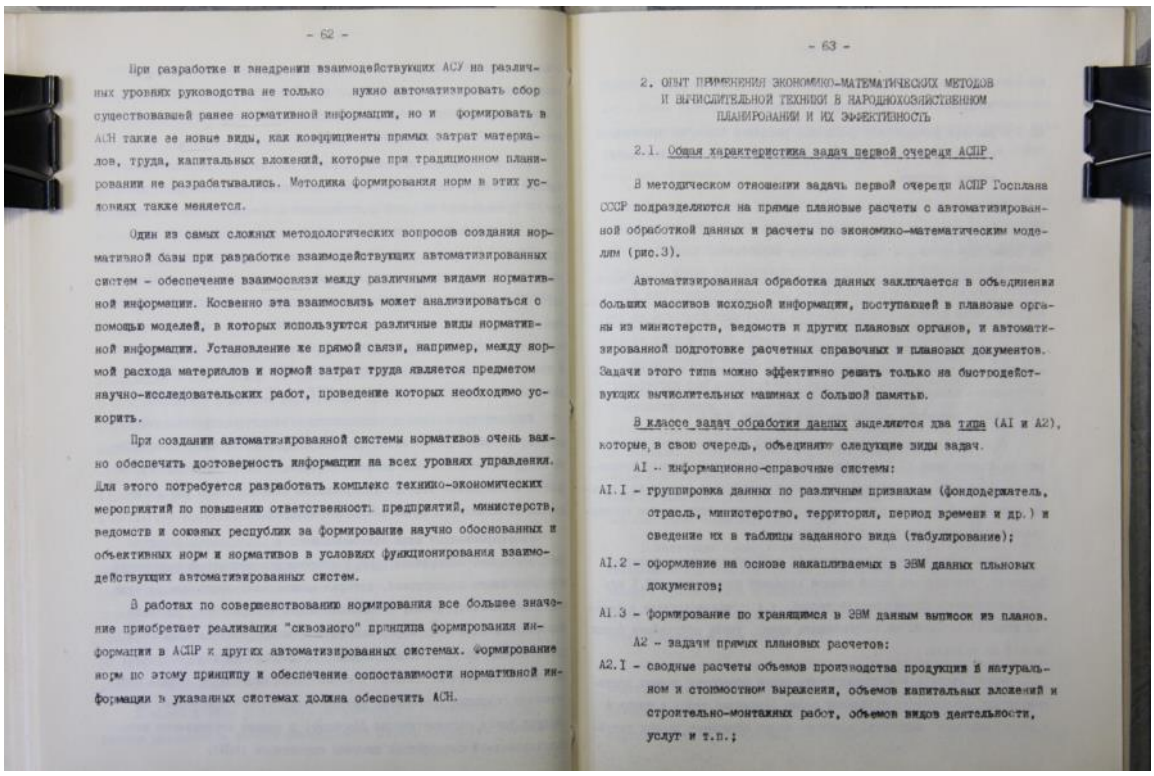


1- фотокамера, 2-осветители, 3 - оригинал
2α - угол поля зрения объектива.
Притемнены зоны, в которых нельзя размещать источники света.

Рисунок 12. Расположение источников света при фотографировании книг

3. **Стекло.** Конечно, если вы один раз случайно где-нибудь в магазине увидели нужную вам книгу, то вы, скорее всего, отфотографируете её как придётся и чем придётся. Но если вы идёте фотографировать книги целенаправленно, и собираетесь оцифровать несколько изданий, то вам очень пригодится кусок обычного стекла. Найдите стекло по размеру чуть больше разворота книги, отмойте его от грязи, края заклейте скотчем, чтобы не порезаться. Слишком большое стекло можно обрезать стеклорезом, это минутная операция, которую вам без проблем сделает любой человек, умеющий держать в руках инструмент (если среди ваших знакомых такого человека нет, можете попросить продавца в хозяйственном магазине). Стекло кладется сверху на книгу, на разворот. Оно прижимает страницы и тем самым решает проблему с искажениями у корешка. Без стекла страницы книги топорщатся, а строчки текста на фотографии выглядят неровными. Сфотографированная страница накрытой стеклом книги выглядит почти так же, как отсканированная. Если стекла нет, можно аккуратно придерживать край страницы пальцем или прижимать другими книгами.

4. **Железные зажимы для бумаг**. В Ленинку проносить стекла нельзя, поэтому совместно с тамошними тетушками была разработана более прогрессивная технология. Два железных зажима для бумаг из канцелярского магазина позволят вам прижать страницы книги по краям к обложке, и они не будут топорщиться и без стекла. Технология в чем-то даже более прогрессивная, так как носить пару зажимов гораздо удобнее, чем кусок стекла, да и между текстом и объективом нет никаких преград (стёкла всё-таки грязнятся), но придётся каждую страницу зажимать отдельно, что снижает скорость работы. Сильное замедление процесса – единственный минус зажимов.



- 62 -

При разработке и внедрении взаимодействующих АСУ на различных уровнях руководства не только нужно автоматизировать сбор существовавшей ранее нормативной информации, но и формировать в АСН такие же новые виды, как коэффициенты прямых затрат материалов, труда, капитальных вложений, которые при традиционном планировании не разрабатывались. Методика формирования норм в этих условиях также меняется.

Один из самых сложных методологических вопросов создания нормативной базы при разработке взаимодействующих автоматизированных систем - обеспечение взаимосвязи между различными видами нормативной информации. Косвенно эта взаимосвязь может анализироваться с помощью моделей, в которых используются различные виды нормативной информации. Установление же прямой связи, например, между нормой расхода материалов и нормой затрат труда является предметом научно-исследовательских работ, проведение которых необходимо ускорить.

При создании автоматизированной системы нормативов очень важно обеспечить достоверность информации на всех уровнях управления. Для этого потребуется разработать комплекс технико-экономических мероприятий по повышению ответственности предприятий, министерств, ведомств и союзных республик за формирование научно обоснованных и объективных норм и нормативов в условиях функционирования взаимодействующих автоматизированных систем.

В работах по совершенствованию нормирования все большее значение приобретает реализация "сквозного" принципа формирования информации в АСПР к других автоматизированных системах. Формирование норм по этому принципу и обеспечение сопоставимости нормативной информации в указанных системах должна обеспечить АСН.

- 63 -

2. ОПЫТ ПРИМЕНЕНИЯ ЭКОНОМИКО-МАТЕМАТИЧЕСКИХ МЕТОДОВ И ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ В НАРОДНОХОЗЯЙСТВЕННОМ ПЛАНИРОВАНИИ И ИХ ЭФФЕКТИВНОСТЬ

2.1. Общая характеристика задач первой очереди АСПР

В методическом отношении задачи первой очереди АСПР Госплана СССР подразделяются на прямые плановые расчеты с автоматизированной обработкой данных и расчеты по экономико-математическим моделям (рис.3).

Автоматизированная обработка данных заключается в объединении больших массивов исходной информации, поступающей в плановые органы из министерств, ведомств и других плановых органов, и автоматизированной подготовке расчетных справочных и плановых документов. Задачи этого типа можно эффективно решать только на быстродействующих вычислительных машинах с большой памятью.

В классе задач обработки данных выделяются два типа (А1 и А2), которые, в свою очередь, объединяют следующие виды задач.

А1 - информационно-справочные системы:

А1.1 - группировка данных по различным признакам (фондодержатель, отрасль, министерство, территория, период времени и др.) и сведение их в таблицы заданного вида (табулирование);

А1.2 - оформление на основе накапливаемых в ЭВМ данных плановых документов;

А1.3 - формирование по хранящимся в ЭВМ данным выписок из планов.

А2 - задачи прямых плановых расчетов:

А2.1 - сводные расчеты объемов производства продукции в натуральном и стоимостном выражении, объемов капитальных вложений и строительно-монтажных работ, объемов видов деятельности, услуг и т.п.;

Рисунок 13. Использование зажимов для бумаг. Разворот книги занимает весь кадр.

5. **Перспективные искажения.** Поверхность стола горизонтальна, а фотоаппарат вы по-любому держите хоть немного под наклоном т.к. не висите над столом, а стоите рядом. Следовательно, дальний край книги будет на кадре меньше, а строчки будут выглядеть как начальные титры фильма "Звёздные войны". Подложите под дальний край стола что-нибудь, чтобы стол наклонился к вам, и была возможность сделать кадр ровно. Если стол прибит к полу, то можно подложить что-нибудь под дальний край книги. Или да, нависайте над ней.

Рисунок 14. Если вы не Джордж Лукас, то перспективные искажения вам ни к чему.

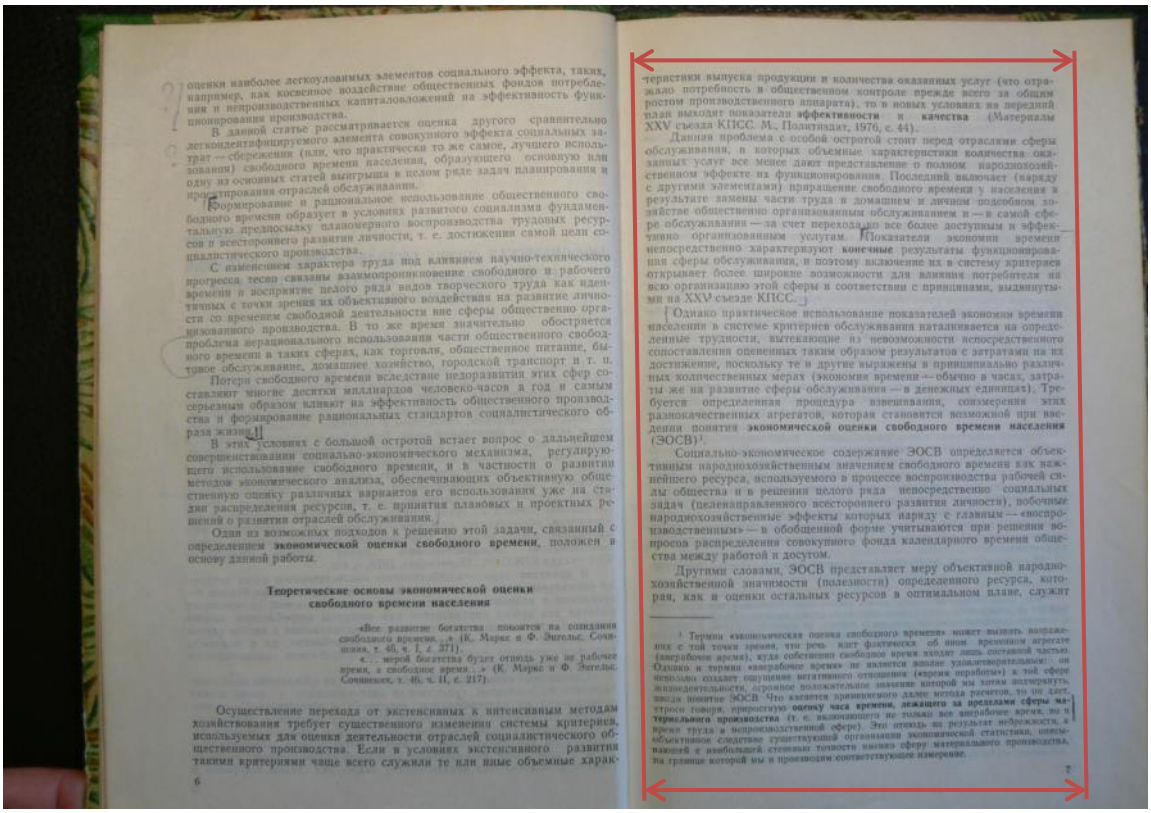


Рисунок 15. А вот так – правильно. Верхний край страницы на снимке имеет такой же размер, как и нижний.

6. **Блики.** Стекло + лампы = блики. Сфотографируйте первый разворот, посмотрите, есть ли блики от ламп. Подвигайте лампы/стол, чтобы их не было. Кстати наклон стола, чтобы не было искажений – ещё и верный способ борьбы с бликами. Если вы фотографируете на улице, то лучше всего, если за вашей спиной будет стена с козырьком, так как небо бликует тоже. Использование зажимов для бумаг вместо стекла позволит автоматически решить проблему бликов.

7. **Одежда.** Если вы выйдете в солнечный день снимать в белой майке, бликовать будет не только небо, но и вы. Одевайте тёмное.
   На этом этапе кому-то из читателей может показаться, что я перегибаю палку. Но если вы посмотрите на рисунок 16, то увидите на обложке книги мой портрет. Потому что когда я её фотографировал, лайфхак про стену за спиной и тёмную одежду я ещё не постиг. Книгу с моим отражением на обложке уже растащили по всему интернету.

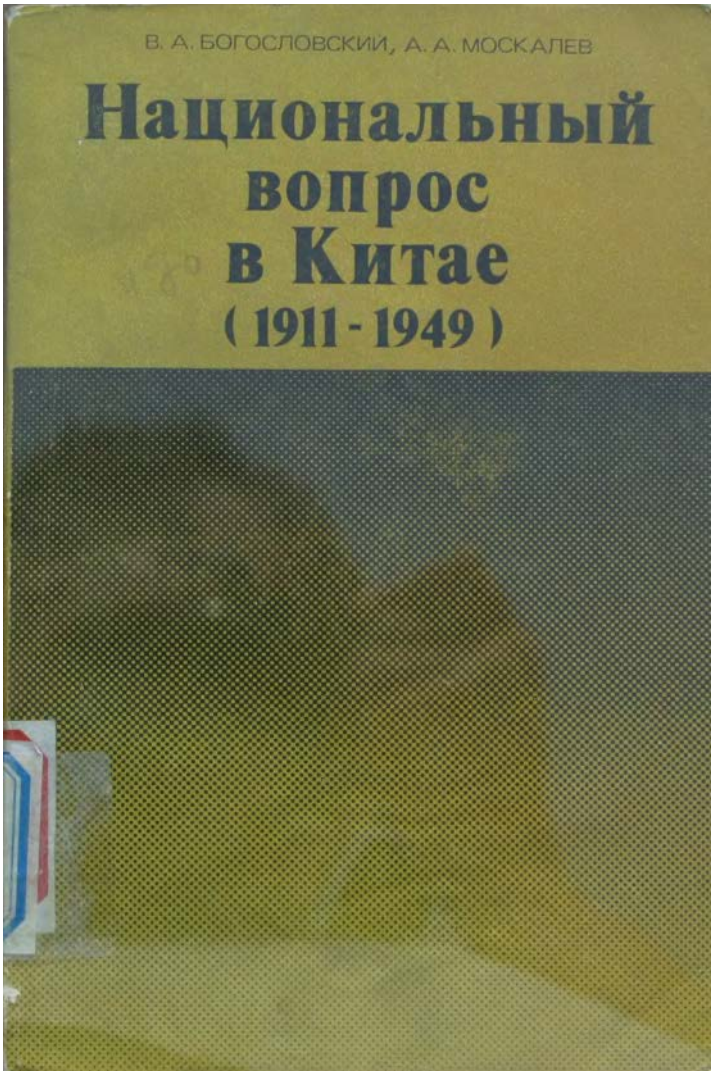


   Рисунок 16. Мое отражение в стекле, которым накрыта фотографируемая книга, появившееся из-за того, что я не позаботился о защите от бликов.
8. **Порядок действий, или «научная организация труда».** Положили книгу, положили сверху стекло, взяли фотоаппарат, оглядели края кадра (что текст влезает, а лишнего нет). Нажали кнопку затвора, отпускаем фотоаппарат (он висит на ремне на шее, так что не падает), одна рука приподнимает стекло, другая рука переворачивает страницу, фотоаппарат в руки, оглядели края кадра… Я успевал отфотографировать до 15 книг за половину светового дня. При определенной сноровке книгу в 300 страниц можно оцифровать за полчаса или даже меньше.

При съемке следует помнить, что всё-таки главное - это резкость, освещенность. С искажениями можно бороться хитрыми программами. С размытыми буквами бороться практически невозможно. Если света не хватает, или не удается ничего сделать с бликами, уберите стекло и включите вспышку на фотоаппарате. Изогнутые, но четкие от вспышки страницы лучше, чем ровные, но размытые. Но некоторые сорта бумаги от вспышки бликуют не меньше, чем стекло.

Напоследок давайте ещё раз разберём все ошибки, которые можно сделать, фотографируя книгу, и которых легко избежать, если просто чуть внимательнее отнестись к процессу.

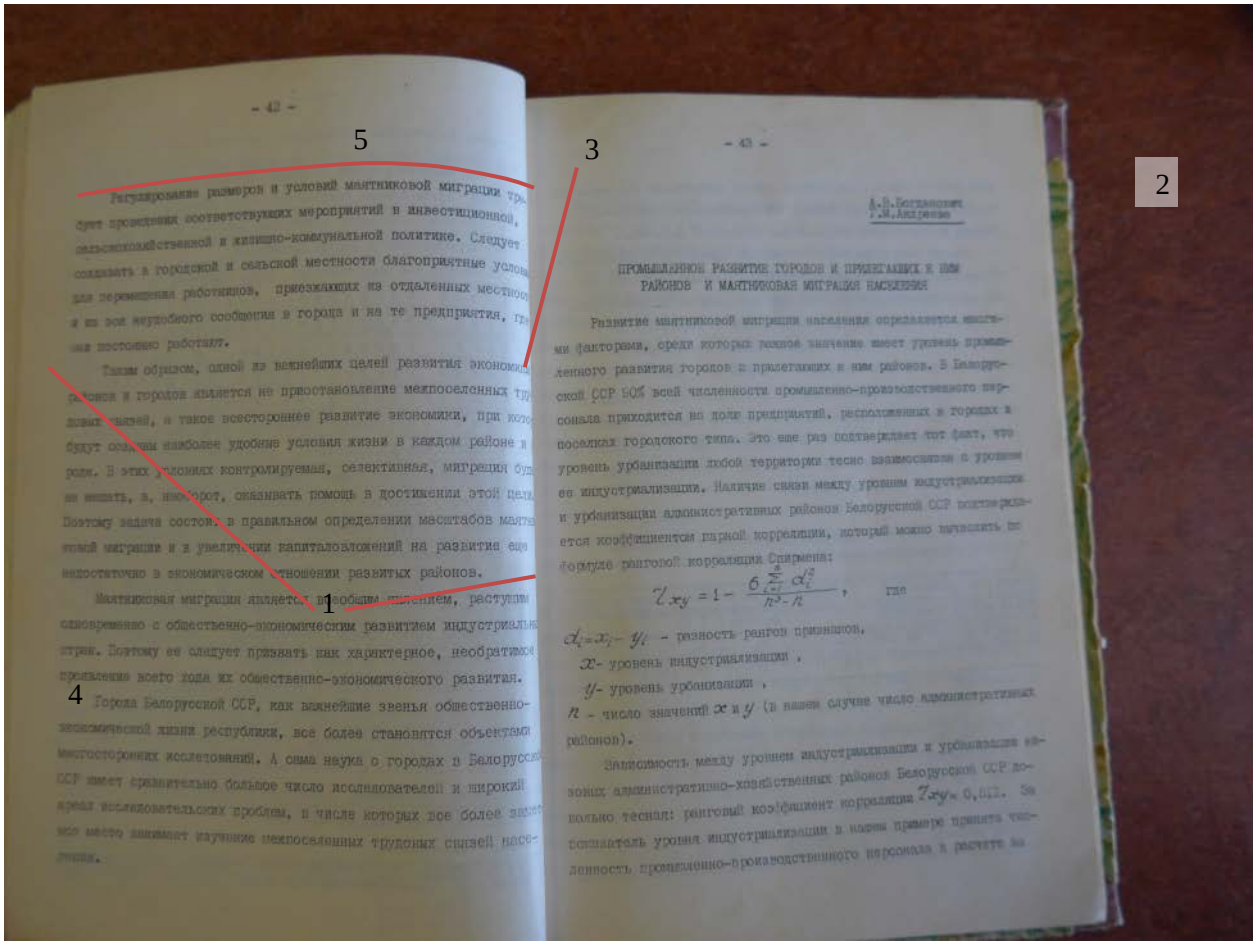


Рисунок 17. Пример неудачной оцифровки страницы. Цифрами обозначены:

1 – перепад освещенности из-за неправильного освещения (видимо, свет из окна справа)

2 – большая часть кадра занята не книгой, а столом

3 – часть текста из-за загибания страниц в кадр не попала. Концы слов обрезаны

4 – буквы у края страниц размыты. Это произошло потому, что в помещении было темно, и фотоаппарат установил маленькую глубину резкости. Фокусировка фотоаппарата произошла по центру кадра, где у корешка страницы были выпуклыми и из-за этого были ближе к фотоаппарату на пару сантиметров. Этих сантиметров хватило, чтобы края книги оказались не в фокусе.

5 – геометрические искажения. Строчки на левой странице имеют волнообразную форму, т.к. страницы не были прижаты ни стеклом, ни зажимами, ни пальцем.

В итоге главную свою функцию такой снимок выполняет: книгу можно прочесть. Если вы – шпион, то миссия выполнена. Но вот собрать из пачки снимков приличную электронную книгу, которую будет удобно читать, из таких сканов не получится. Кроме того, такой текст, скорее всего, файнридер распознает со множеством ошибок. Проще говоря, если вы сами небрежно сфотографируете книгу, потом сами же с ней намучаетесь. Всегда помните, что чем выше качество скана, тем проще его потом обрабатывать.

Ну а теперь перейдём к обработке.

Часть 2. Обработка сканов и кодирование книг

Подготовка сканов

Для обработки сканов вам потребуются следующие программы (в порядке использования):

1. **XnView** (её использование было освещено в предыдущей части).
2. **Adobe Photoshop** с дополнительно установленными плагинами пакета **ScanTools** и **Sattva descreen**;
3. **Book restorer** (опционально, помогает улучшать плохие сканы, но работает крайне нестабильно и может быть заменен **scan tailor**'ом);
4. **Scan tailor featured** (обратите внимание, что нужен именно featured, а не просто Scan tailor);
5. Djvu small;
6. Djvu imager;
7. **Abbyy Fine reader версии 11 и выше**;
8. FR11 DjVu Text Layer Crutch;
9. **Adobe PDF creator** (для кодирования в PDF)
10. PDF&DJVU Bookmarker.

Из этого списка платными являются только фотошоп и файнридер. Остальные программы свободно скачиваются в интернете.

Этап обработки начинается, когда у вас есть пакет «сырых» сканов в формате Tiff в градациях серого. Преобразовать сканы в нужный формат можно с помощью программы XnView.

Обработка сканов начинается в фотошопе, для чего сам фотошоп надо подготовить установкой в него дополнительных плагинов и записью действий для пакетной обработки. Это нужно будет сделать только один раз, не пугайтесь.

Я очень советую вам установить плагин **Sattva Descreen**. Он платный, но стоит копейки, и на мой взгляд свои деньги полностью отрабатывает. На официальном сайте подробно описано, зачем он нужен, и как он работает http://www.descreen.net/rus/soft/descreen/descreen.htm Там же есть подробная инструкция по установке и использованию. Обратите внимание, что для разных версий фотошопа есть разные плагины!

Вкратце - он аккуратно удаляет тот самый типографский растр (точки), из которого состоят все напечатанные типографским способом изображения в книгах. Именно для использования этого плагина надо сканировать страницы с изображениями с разрешением 600 ДПИ. В этом случае каждая точка на скане становится видна, и на них можно «натравить» этот плагин.

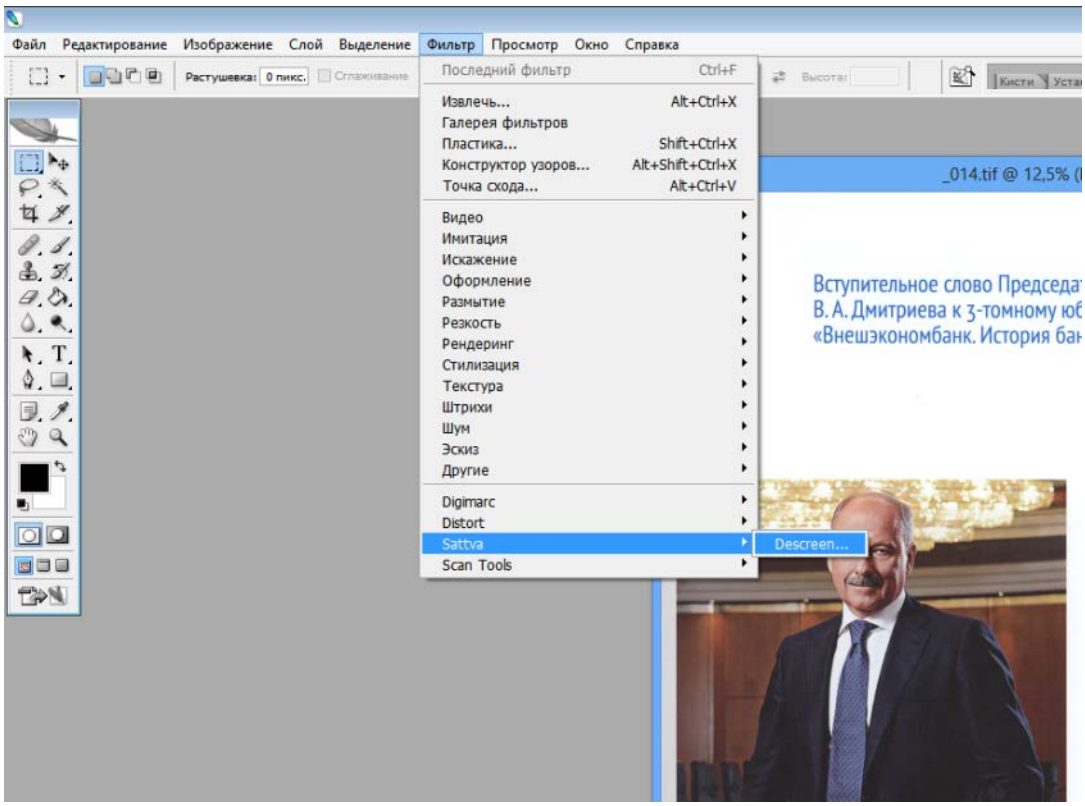


Рисунок 18. Запуск плагина Descreen после установки. Также виден установленный пакет плагинов Scan tools.

Удалять растр с изображений надо в первую очередь, до любых других преобразований! Иначе на изображениях появятся «артефакты» (см. рис. 4) Исключением являются повороты на угол, кратный 90 градусам. Поворачивать на 90/180 градусов можно и до удаления растра.

Выделяем в фотошопе нужную нам область страницы, жмем фильтр – sattva descreen, затем как правило программа сама определяет все параметры (растр и угол растра, угол нужен только для цветных изображений) (если нет, выставляем руками на глаз), жмем ОК, получаем изображение, размытое ровно настолько, чтобы точек не было видно, но при этом мелкие детали не пропадали.

Плагин не умеет самостоятельно определять место на странице, на котором есть изображения! При запуске он показывает вам увеличенный кусочек страницы, и если в этом кусочке изображения нет, то плагин может не сработать (т.е. не определить частоту растра и угол розеточного муара). В этом случае перетащите увеличенную область на то место страницы, где есть рисунок.

Плагин **descreen** применяется ко всей странице, если вы вручную не выделили область с рисунком, он размоет не только рисунки, но и текст. Если вы хотите этого избежать, но вам лень выделять вручную каждый рисунок, можете попробовать укротить плагин **Picture mask** из набора **Scan tools**, речь о котором пойдет ниже. Он по идее должен как раз самостоятельно определять области с рисунками и самостоятельно их выделять, чтобы потом **descreen** применялся только к выделенным участкам, но у меня Picture mask часто ошибался, поэтому я его не использую. Благо что во взрослых книгах рисунков не так много, а детские наоборот такие цветастые и с такими большими буквами, что там можно смело применять **descreen** ко всей странице целиком.

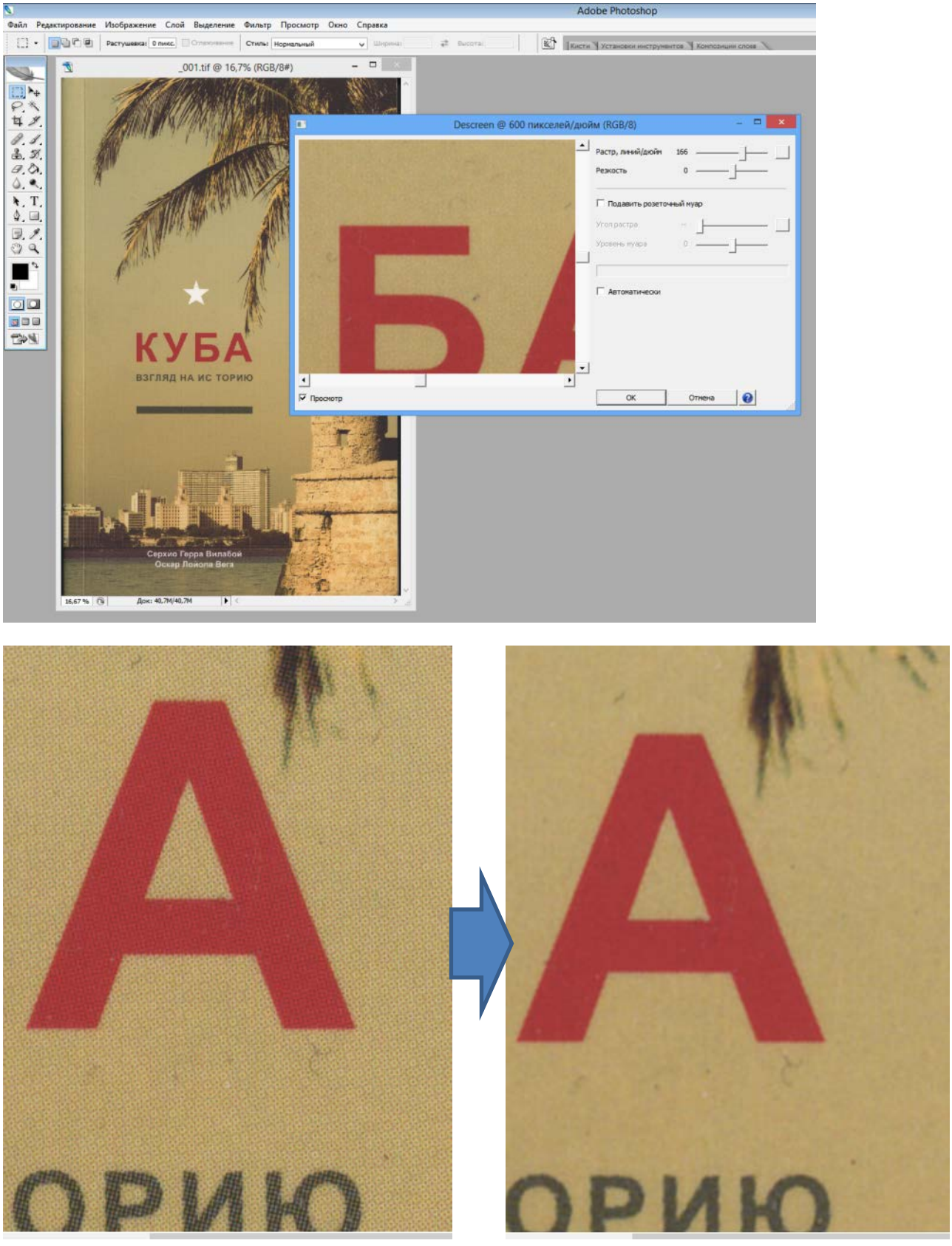


Рисунок 19. Результат работы фильтра descreen.

Пакет плагинов **ScanTools** изначально был доступен на сайте http://abab.front.ru/ScanTools/ScanTools.ZIP но после того, как сайт испустил дух, распространяется «из рук в руки». Скачайте его отсюда https://yadi.sk/d/2rGmzmoFDPCNb.

Распакуйте архив, и скопируйте файлы оттуда в папку в фотошопе, где находятся файлы с таким же расширением (у меня это папка C:\Profram files\Adobe\Adobe photoshop CS2\Внешние модули\фильтры).

После этого у вас на вкладке фильтры появится новый блок фильтров Scan tools. Краткая инструкция по работе с ними есть в том же архиве. Вам в первую очередь будет нужен фильтр **Background clear** и реже – фильтр **Despecle.** Остальные фильтры из набора я почти не использую.

Фильтр **Background clear**, как ясно из названия, радикально высветляет фон, не трогая при этом буквы и изображения. Единственная проблема – если какие-то буквы/изображения обрезаны по краю кадра, фильтр высветлит и их тоже. Бороться с этим можно, например, нарисовав кистью светлую полоску, которая отделит нужный элемент от края страницы.

Если при попытке применить фильтр на серое (не цветное) изображение фотошоп ругается, значит оно недостаточно «не цветное». Именно для удобной работы с этим фильтром мы и использовали XnView, чтобы сделать изображение воистину не цветным (см. первую часть инструкции).

В результате ваши серые сканы станут белыми. После работы фильтра можно по желанию применить автоматическую (Alt+Ctrl+L) или ручную тоновую коррекцию (инструмент уровни/levels, вызывается командой Ctrl+L).

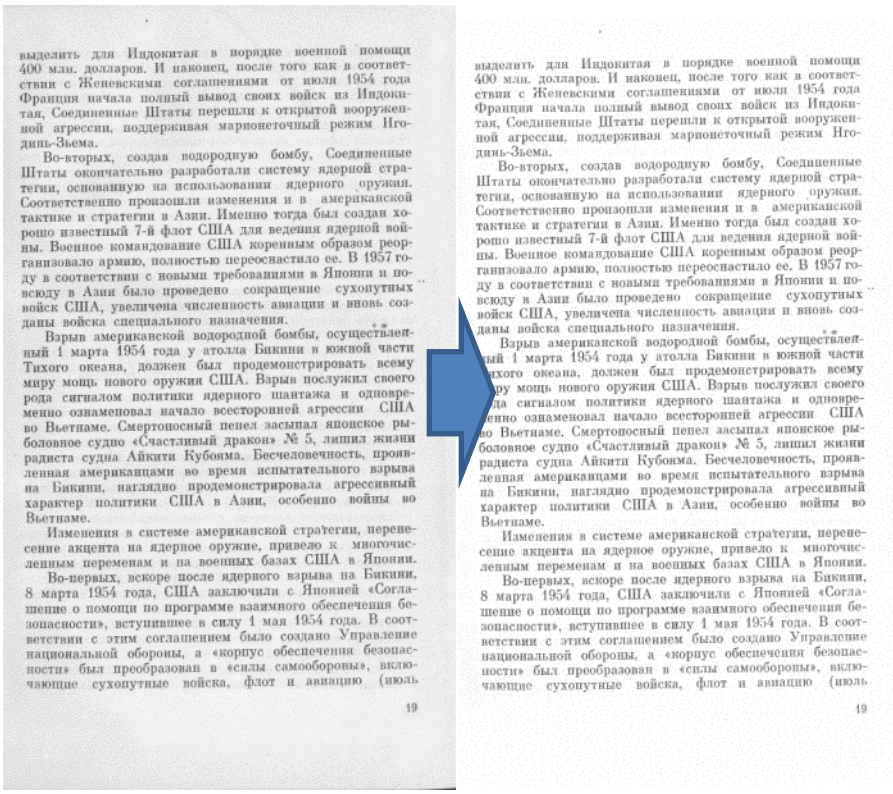


Рисунок 20. Результат работы фильтра Background clear.

Испробуйте этот фильтр на нетипичных страницах:

1. с близко или вообще вплотную к краям расположенным текстом
2. с ярко выраженными (широкими, интенсивными) темными тенями у краев
3. с участками с низким контрастом, когда яркость фона и текста близки, особенно темный фон и темный текст
4. со страницами, где мало текста -- 2-3 слова или одна-две строки.

Вы должны понимать, в каких случаях можно использовать его «не глядя» (сейчас дойдём до пакетной обработки), а когда лучше работать с каждым сканом отдельно.

Страницы с иллюстрациями лучше обрабатывать вручную. Иногда фильтр ошибается, и высветляет и их тоже.

Я при обработке книг перемещаю страницы с рисунками в отдельную папку и обрабатываю их вручную: выделяю область с рисунком, применяю к ней фильтр **descreen**, потом инвертирую выделенную область (команда ctrl+shift+ I) и применяю фильтр **background clear**. Фон всей страницы, кроме рисунка, выбеляется. Потом я инвертирую выделенную область обратно (чтобы опять оказался выделен рисунок) и вручную подручиваю уровни (уровни можно вызвать нажатием клавиш ctrl+L) чтобы сделать рисунок чуточку контрастнее. Белый фон остальной страницы при этом служит ориентиром.

Фильтр **Despecle** убирает мелкий мусор со сканов. Его надо применять, когда бумага книги столь плохая, что опилки и волокна целлюлозы явственно видны, либо когда на листе кроме букв есть ещё мелкие черточки краски. То и другое часто встречается в довоенных книгах. В архиве есть аннотация, если кратко – в фильтре 3 параметра:

Первый - размер в пикселах того мусора, который убирается на всей площади страницы.

Второй - размер зоны сохранения вокруг крупных объектов (букв), внутри которой не работает третья настройка.

Третий - размер в пикселах того мусора, который дополнительно убирается на той части страницы, которая выходит за пределы зоны сохранения.

К примеру, если задать там 3, 7, 10, то сначала программа удалит все точки диаметром 3 и менее пикселов, потом обведет оставшиеся объекты полосой шириной 7 пикселов, и после удалит всё за пределами этой полосы, что меньше 10 пикселов.

Будьте аккуратны, программа любит «есть» точки в конце предложений, над английскими буквами i (которые часто встречаются в формулах), точки, которыми иногда отграничивают поля в таблицах, и другие полезные элементы текста. Если буквы плохо пропечатаны, фильтр сожрёт и фрагменты букв. В сносках шрифт текста меньше, поэтому текст в сносках может пострадать особенно сильно. Поэтому обязательно вручную поиграйте с настройками указанных параметров, а в «легких» случаях вообще не используйте его. От греха.

Освоив фильтры, надо один (один!) раз записать акцию (действие) для автоматической пакетной обработки всей нашей папки со сканами.

Инструкций, как это сделать, в интернете много, вкратце – последовательность следующая: «создать действие – название действия - записать (начинается запись, дальше программа записывает все наши действия) - открыть файл - фильтр - сохранить как - закрыть файл - стоп).

Обратите внимание, что записаны будут именно те параметры фильтров, которые там стояли, когда мы делали запись. Если нужно обработать файлы с другими настройками фильтров, процедуру записи придется повторить.

Записать можно любую комбинацию действий. Вы можете обрезать изображение, высветлить его, убрать растр, повернуть на определенный угол, повысить резкость, применить тоновую коррекцию, сохранить в другом формате, и всё это будет записано как одно действие. Разумеется, в автоматическом режиме можно применять и фильтр **descreen**, только надо помнить, что он будет

применяться именно с теми параметрами размытия, которые были выставлены при записи, а не с теми, которые будут определены исходя из ваших изображений. Поскольку растр в разных книгах разный, я или обрабатываю каждую страницу с изображениями вручную (обычно их немного), или перед пакетной обработкой записываю новое действие с фильтром **descreen**, применяя его к странице из той же книги, которую потом буду обрабатывать «в пакете».

Теперь фотошоп может самостоятельно обработать все наши файлы в автоматическом режиме. На вкладке «файл» надо выбрать «автоматизация», а потом – «пакетная обработка». Там выбираем только что записанную акцию (набор действий), указываем источник (папку со сканами) и назначение (создаем другую папку, пустую, куда будет сохраняться результат).

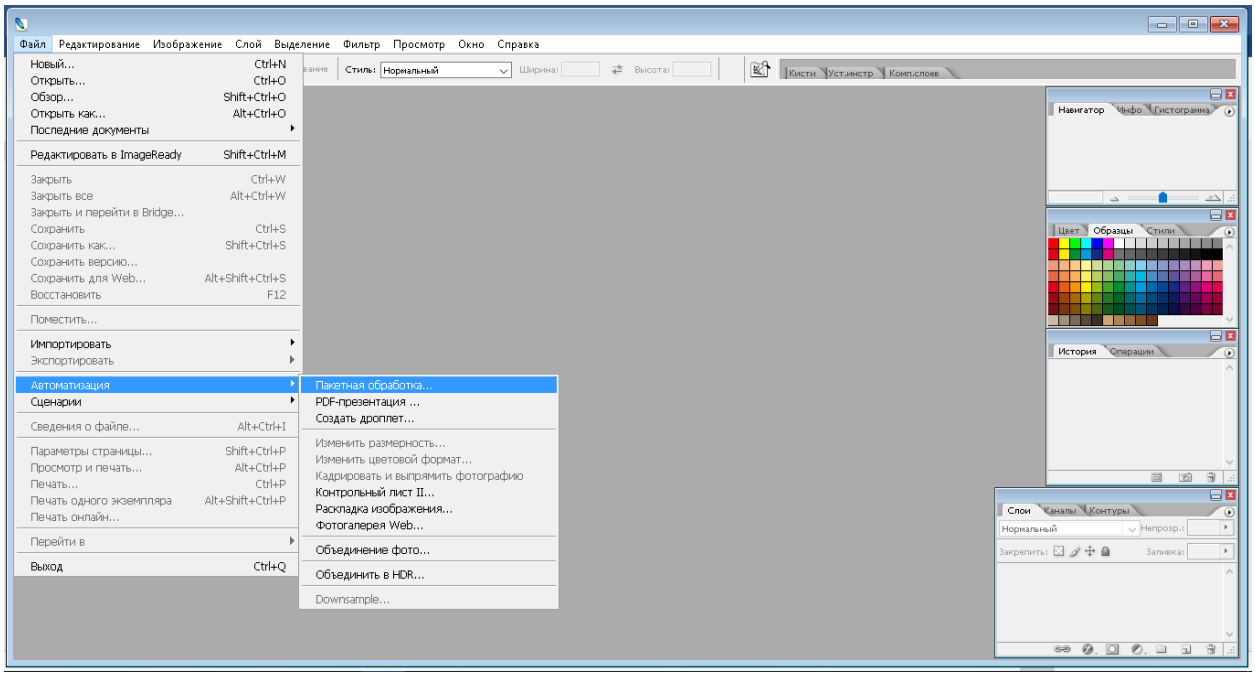


Рисунок 21. Запуск пакетной обработки.

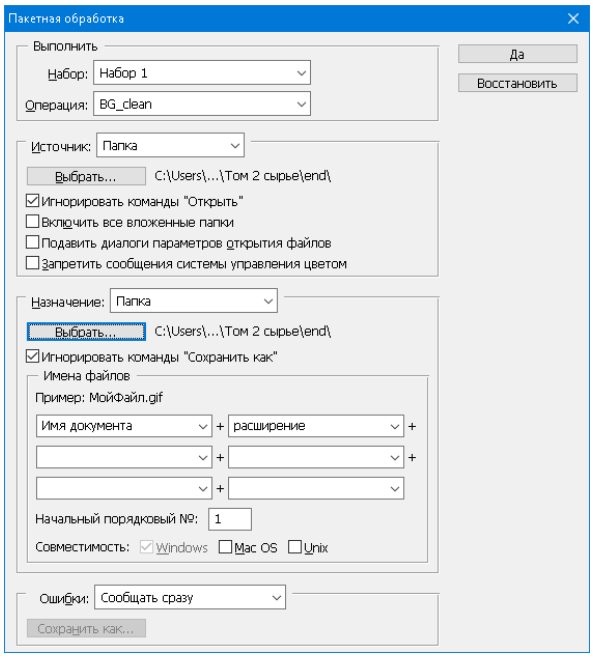


Рисунок 22. Выбор действия и папок с исходными файлами и местом сохранения результатов для пакетной обработки.

Я очень советую создавать отдельные подпапки для каждой следующей операции. Пусть лучше у вас вырастет дерево папок end – end2 – end3 и так далее, чем вы тем же самым неудачно настроенным фильтром **despecle** попортите исходные сканы.

Теперь спокойно ждём, пока нужным фильтром будут обработаны все сканы.

Разумеется, возможности фотошопа гораздо шире. Можно в ручном режиме, работая с отдельными сканами, удалять печати библиотек, цветные подчеркивания (когда кто-то читал книгу с ручкой в руке), освежать цвета обложки, редактировать фотографии… Это простор для творчества, и здесь я вам не советчик. Каждая задача индивидуальна.

По умолчанию все сканы требуют прогонки через фильтр **Background clear**, а сканы с изображениями – через фильтр **Descreen**. Остальное опционально.

Программа **Book restorer** будет вам нужна в том случае, если ваши сканы ужасны. Она позволяет распрямлять кривые строчки и выравнивать освещенность, но сама по себе часто глючит и вылетает. Аккуратное сканирование избавит вас от необходимости работы с ней, поэтому подробно я на ее использовании не останавливаюсь. Инструкции, разумеется, есть в интернете. К сожалению, сама программа работает крайне нестабильно. Строки можно выпрямлять и **Scan Tailor**'ом, эта опция доступна на последнем этапе обработки в нём.

Программа **Scan tailor featured** является основным инструментом подготовки сканов. Она проста, надежна, интуитивно понятна, содержит всё необходимое и не содержит ничего лишнего. В неё надо загрузить наши сканы, прошедшие через нужные фильтры (указать папку с ними при открытии программы).

Весь процесс обработки разбит на 6 шагов: Исправление ориентации, разрезка страниц, компенсация наклона, выделение полезной области, создание полей, вывод готовых изображений. В большинстве случаев программа отлично работает автоматически, но каждый шаг можно скорректировать вручную.

Комментариев требуют, пожалуй, только три последних шага.

Полезную область программа определяет, анализируя темные элементы на странице и вписывая их все в прямоугольник. Если на странице помимо текста есть ещё какая-нибудь грязь, или на скане виден корешок книги или ваши пальцы, то полезную область (область текста) программа определит неправильно. После того, как она автоматически определит полезную область на всех страницах, надо отсортировать сканы по возрастающей высоте (высоте полезной области), и просмотреть самые большие и самые маленькие (начало и конец списка). При необходимости поправить границы полезной области вручную. Затем аналогичным образом надо отсортировать сканы по возрастающей ширине и повторить процедуру.

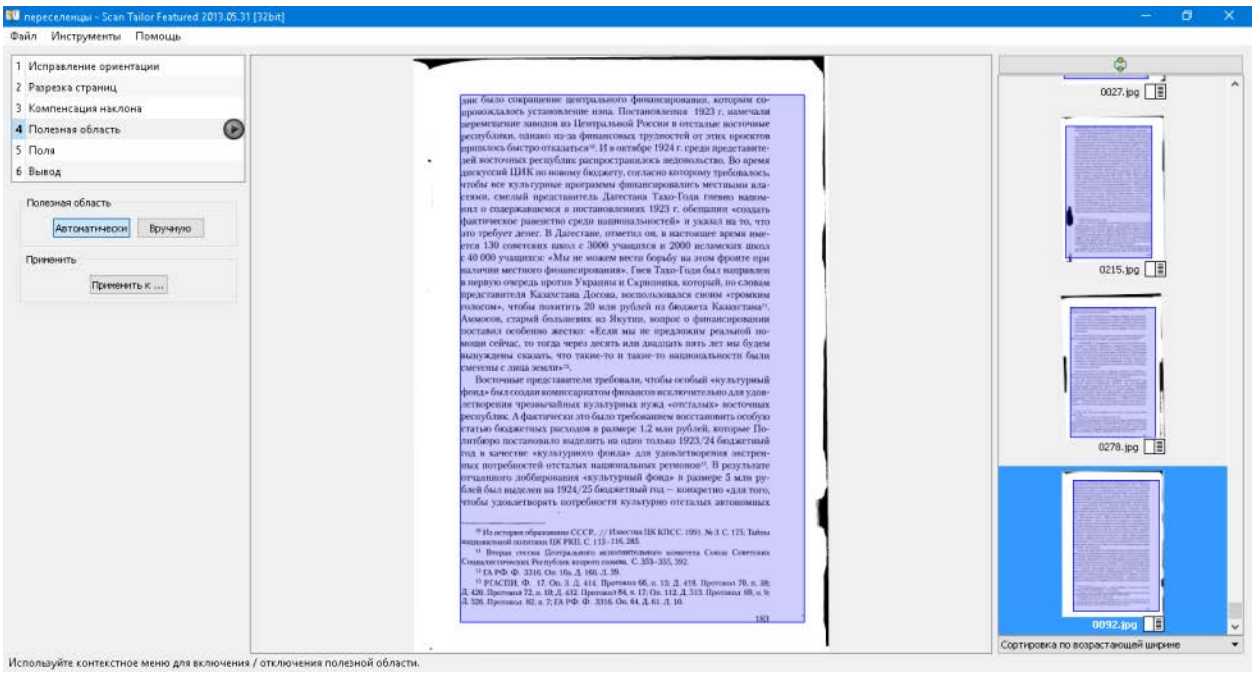


Рисунок 23. На рисунке показан этап определения полезной области. Программа ошиблась, границы полезной области справа от текста больше, чем нужно. Придется поправить вручную, подвинув границу синего прямоугольника так, чтобы он не выходил за границы текста.

Поля я обычно задаю шириной 3 мм. Поля обложки я задаю равными 0 мм, и снимаю флажок «выровнять размеры с другими страницами». Перед этим надо выделить как полезную область всю обложку. В результате у меня обложки имеют свой собственный размер и не имеют полей, а все текстовые страницы имеют одинаковый размер, установленный по наибольшей из них.

Если поля получаются какие-то аномально большие, значит вы проглядели на какой-то из страниц слишком большую полезную область. Второе объяснение – вы проглядели какую-то широкую страницу (например, вкладку с картой), и все поля оказались не меньше, чем эта самая большая страница.

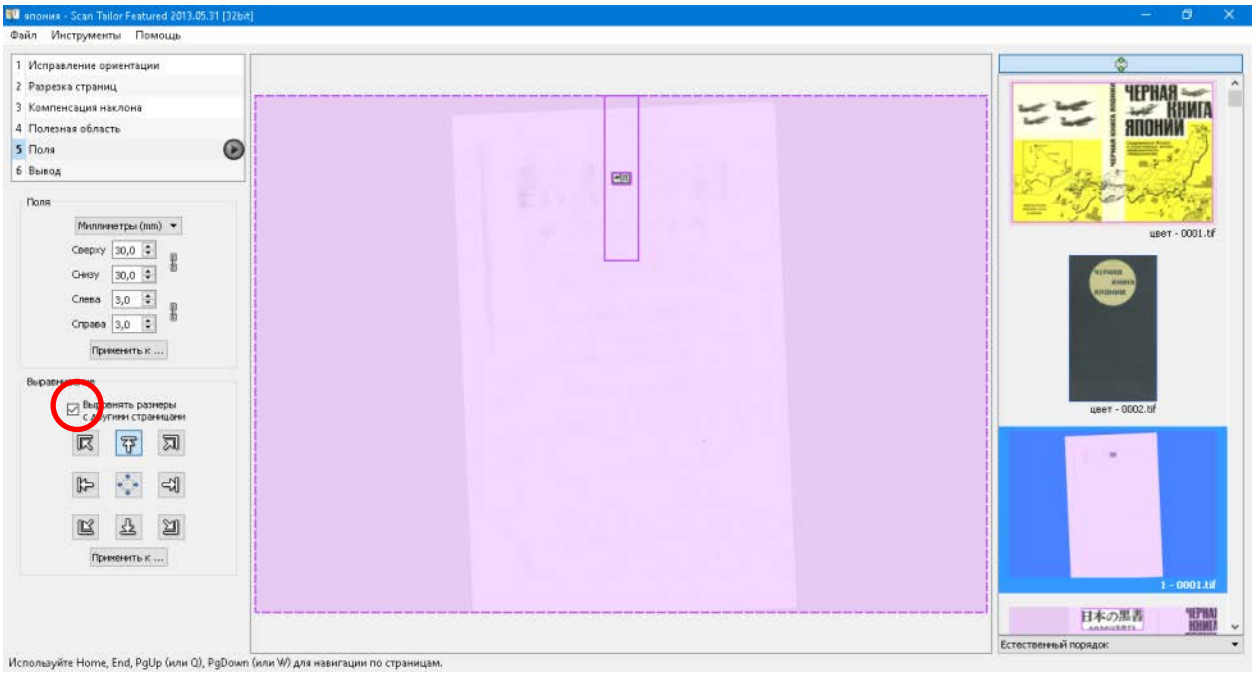


Рисунок 24. Первой страницей книги идет широкая суперобложка (видна как миниатюра в верхнем правом углу). При задании полей для нее забыли снять галочку «выровнять размеры с другими страницами», и поля всех страниц оказались очень широкими. Маленьким прямоугольником в центре показаны «собственные» поля выбранной страницы.

Иногда текст расположен в центре (название книги) или внизу (выходные данные и тираж) страницы. На этапе задавания ширины полей можно выровнять текст по середине или по любому краю страницы. Удобно, установив ширину полей, отсортировать страницы по возрастающей высоте, и просмотреть самые маленькие. Как правило это и есть страницы с тем текстом, который в бумажной книге был напечатан не сверху страницы.

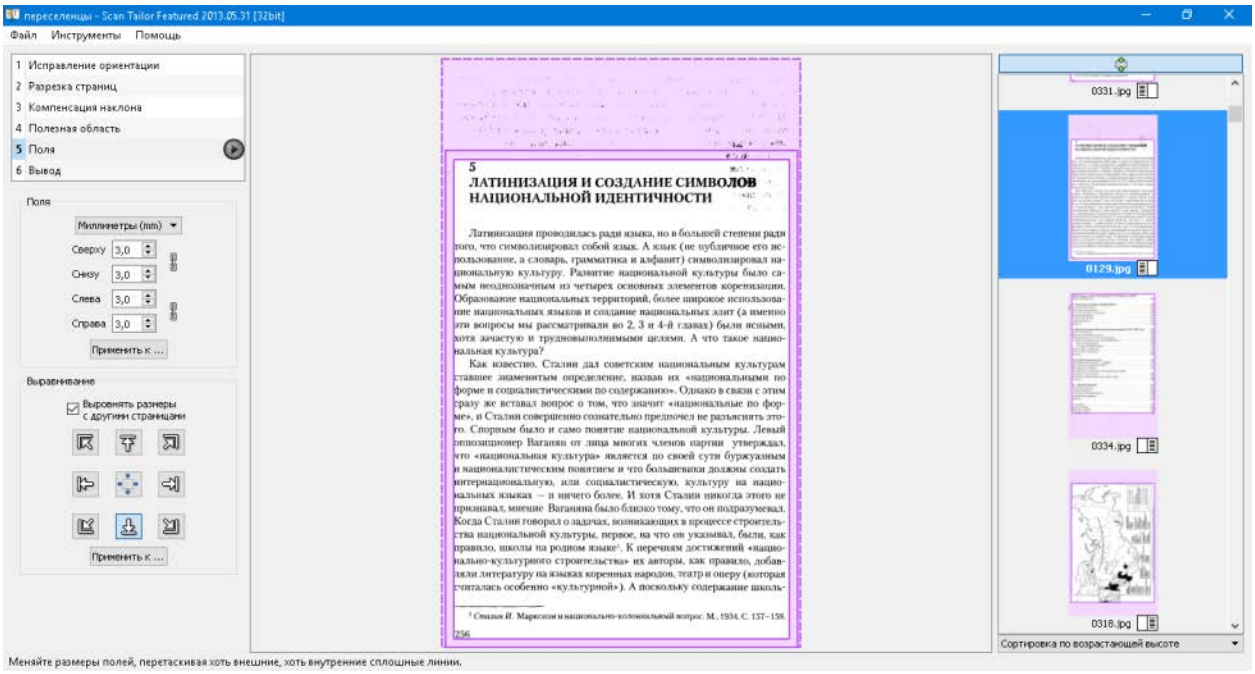


Рисунок 25. На рисунке идет работа со страницей, с которой начинается новая глава книги. Применено выравнивание по нижнему краю, чтобы текст был там же, где в бумажном оригинале – внизу.

Не пренебрегайте процедурой выравнивания. Во-первых, это просто красиво. Если редакторы книги сделали текст посередине или внизу страницы, то, наверное, это было сделано не зря. Если у вас он уедет вверх (а по умолчанию так и будет), то будет немного грустно. Это те самые мелочи, которые отделяют хорошо сделанную (вашу) работу от посредственной. Во-вторых, в процессе выравнивания вы сможете заметить неверно определенные полезные области, которые не заметили на предыдущем этапе.

Если на этом месте вы уже перестали понимать смысл прочитанного, то установите, наконец, **Scan tailor featured** и попробуйте обработать любую книгу, хотя бы десяток страниц. Всё сразу станет понятно, уверяю вас.

Самый ответственный этап работы программы – вывод. Вывод возможен в трех вариантах: черно-белый, цветной/серый, смешанный.

При выборе черно-белого варианта происходит бинаризация. Все точки темнее определенного порога бинаризации становятся абсолютно черными, а светлее – абсолютно белыми. Порог бинаризации можно менять, делая текст «жирнее» или «тоньше». Этот режим удобен для книг без картинок.

Вариант «цветной/серый» фактически только выравнивает и подрезает страницы, но никак их не преобразует. Я использую его, если по каким-то причинам надо сохранить фон книги, или если сканы не очень хорошие, и бинаризация дает плохой результат.

Иногда вариант «цветной/серый» приходится выбирать, если по каким-то причинам сканы требуют дополнительной обработки после нарезки. К примеру, если вы фотографировали книгу фотоаппаратом с рук, то страницы книги будут иметь разный размер. Где-то вы держали фотоаппарат чуть ближе, а где-то – чуть дальше. В этом случае после разрезки страниц, выделения полезной области и создания полей имеет смысл снять галочку «выровнять с другими страницами» и сохранить результаты в этот самом «цветном/сером» режиме. Получим набор нарезанных и выровненных страниц разного размера. Затем в фотошопе надо записать действие по увеличению размера изображения, до величины, соответствующей самой большой странице, и прогнать пакет сканов через это действие. Результатом станет набор страниц одного размера. После этого их можно либо опять загрузить в **scan tailor** и бинаризовать в нём, либо воспользоваться для бинаризации одним из фильтров того пакета, который я выше предлагал вам установить. Помимо **background clear** и **despecle** он включает еще два инструмента для бинаризации с настраиваемым порогом бинаризации (т.е. настраиваемой границей того уровня освещенности, выше которого после бинаризации всё становится абсолютно белым, а ниже – абсолютно черным). Это фильтры **otsu binarization** и **threshold binarization**.

Иногда страницы после scan tailor надо не увеличивать, подгоняя под единый размер, а, допустим, повышать резкость или контрастность изображения. Во всех случаях, когда требуется последующая обработка, вывод следует проводить в режиме «Цветной/серый».

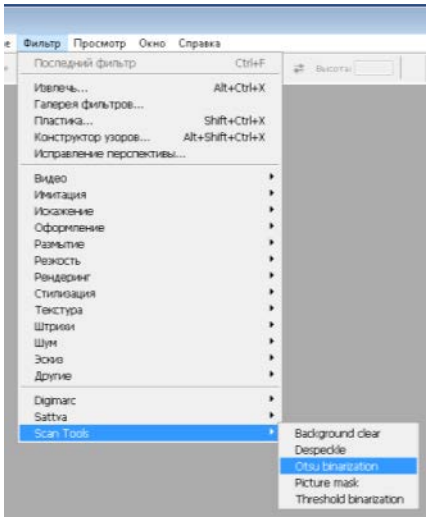


Рисунок 26. Фильтры (плагины) из пакета **Scan tools** для бинаризации изображений в фотошопе. Могут применяться уже после **Scan tailor**'а, если по каким-то причинам вывод был произведен в режиме «цветной/серый».

Самый интересный вариант вывода – «смешанный». Программа сама определяет картинки на страницах, картинки не трогает, а области с текстом бинаризует. Как обычно, если программа ошиблась, область картинок можно задать вручную. Смешанный вывод необходим для последующей вставки рисунков в книгу в хорошем качестве.

Параметры вывода можно установить отдельно для каждой страницы. Обложки и страницы с картинками я обычно вывожу в режиме «смешанный», а страницы текста – в черно-белом режиме.

В самом смешанном режиме есть три варианта распознавания изображений: форма картинок свободная, обведенная и квадро. При выборе свободной формы программа будет выделять

картинки по контуру, при выборе обведенной или квадро – вписывать в прямоугольник. Выделение по контуру может промахиваться (например, часто программа не считает картинкой часть фотографии, на которой изображено небо), а выделение «обведенное» или «квадро» захватывает лишний фон.

Если выбрать вкладку «зоны картинок», программа покажет какие части страницы она отнесла к картинкам. На рисунке ниже программа выделила по контуру фотографии самолётов (подсвечиваются синим), но не поняла, что желты фон карты Японии – тоже картинка. Поэтому пришлось вручную нарисовать еще два многоугольника (красные линии, синий фон) чтобы никакие картинки не потерялись.

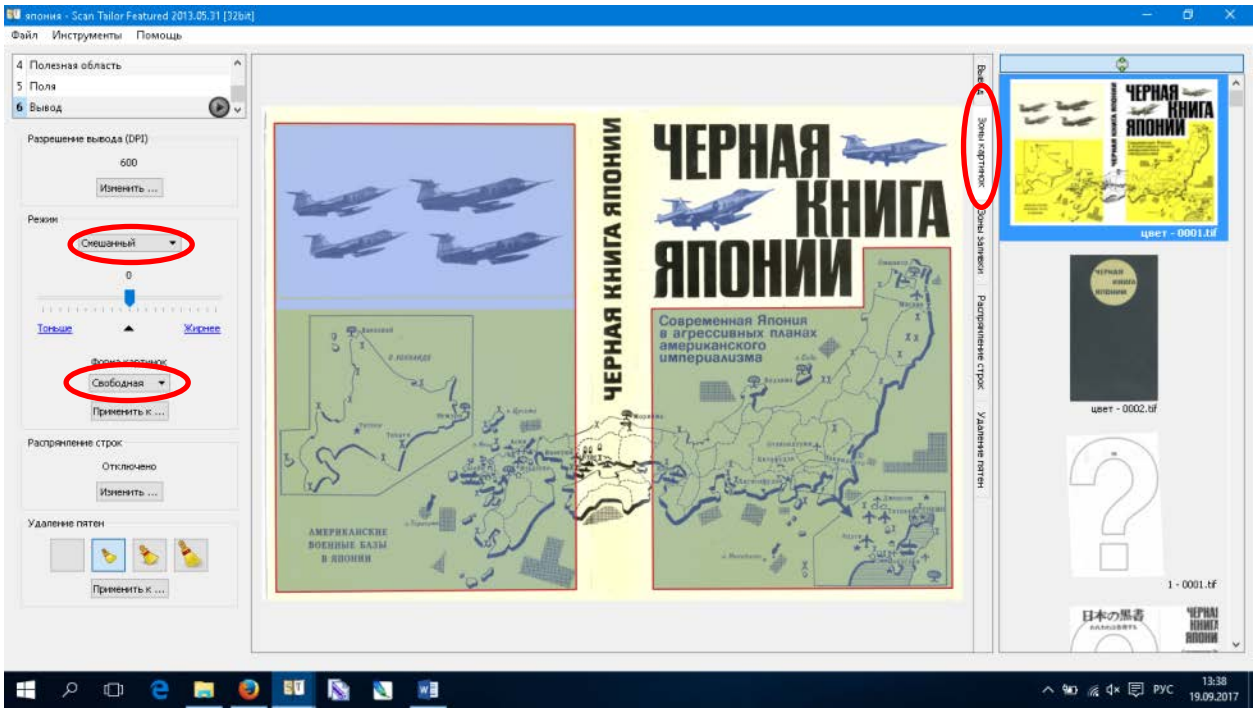


Рисунок 27. Этап вывода, вывод в режиме «смешанный», форма картинок «свободная», активна вкладка «зоны картинок».

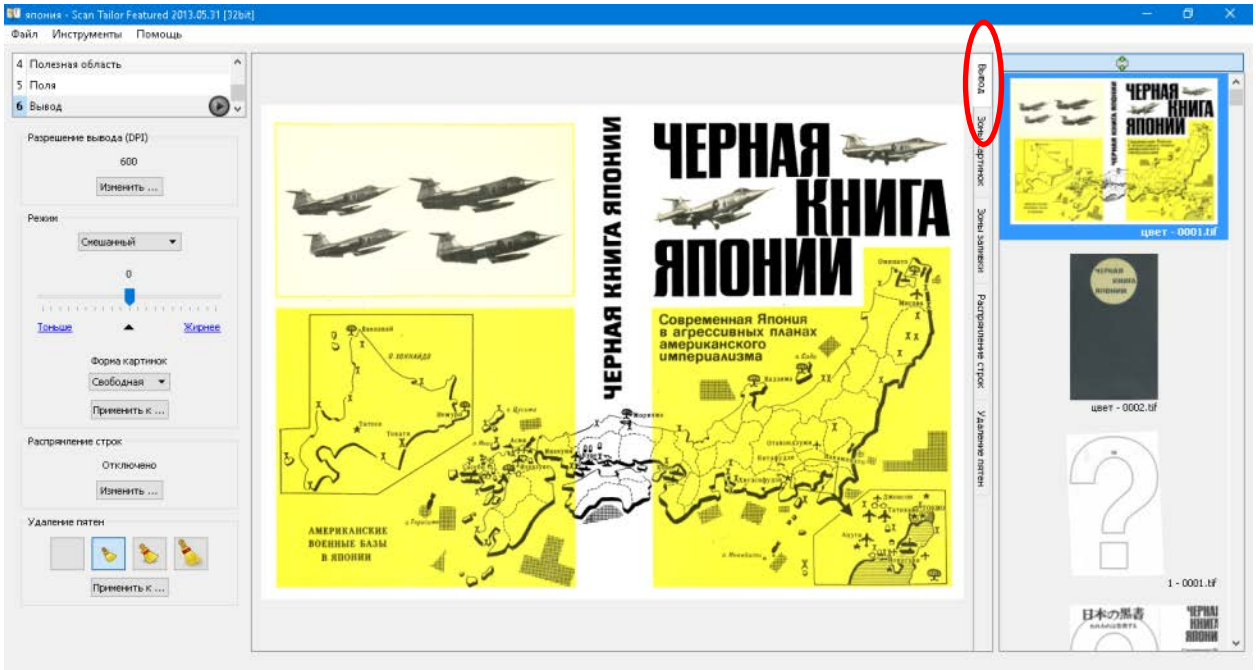


Рисунок 28. Этап вывода, активна вкладка «вывод», показан результат обработки. Области за пределами зон картинок стали черно-белыми, картинки сохранены.

Этап вывода может занимать довольно много времени. Программа фактически создает новые изображения на основе наших сканов. В любой момент процесс можно остановить, сохранить проект, а потом продолжить. Вообще проект лучше сохранять сразу же, во избежание.

После окончания вывода нужно сделать еще одно последнее действие, ради которого я так настойчиво просил устанавливать именно **Scan tailor featured,** а не просто **Scan tailor.** В нем есть опция «экспорт разделенных сканов», и ей нужно воспользоваться. Она создаст в папке «out» с результатами новую подпапку «export», а в ней – подпапки «1» и «2». В подпапке «1» будут только черно-белые страницы. В подпапке «2» – только изображения.

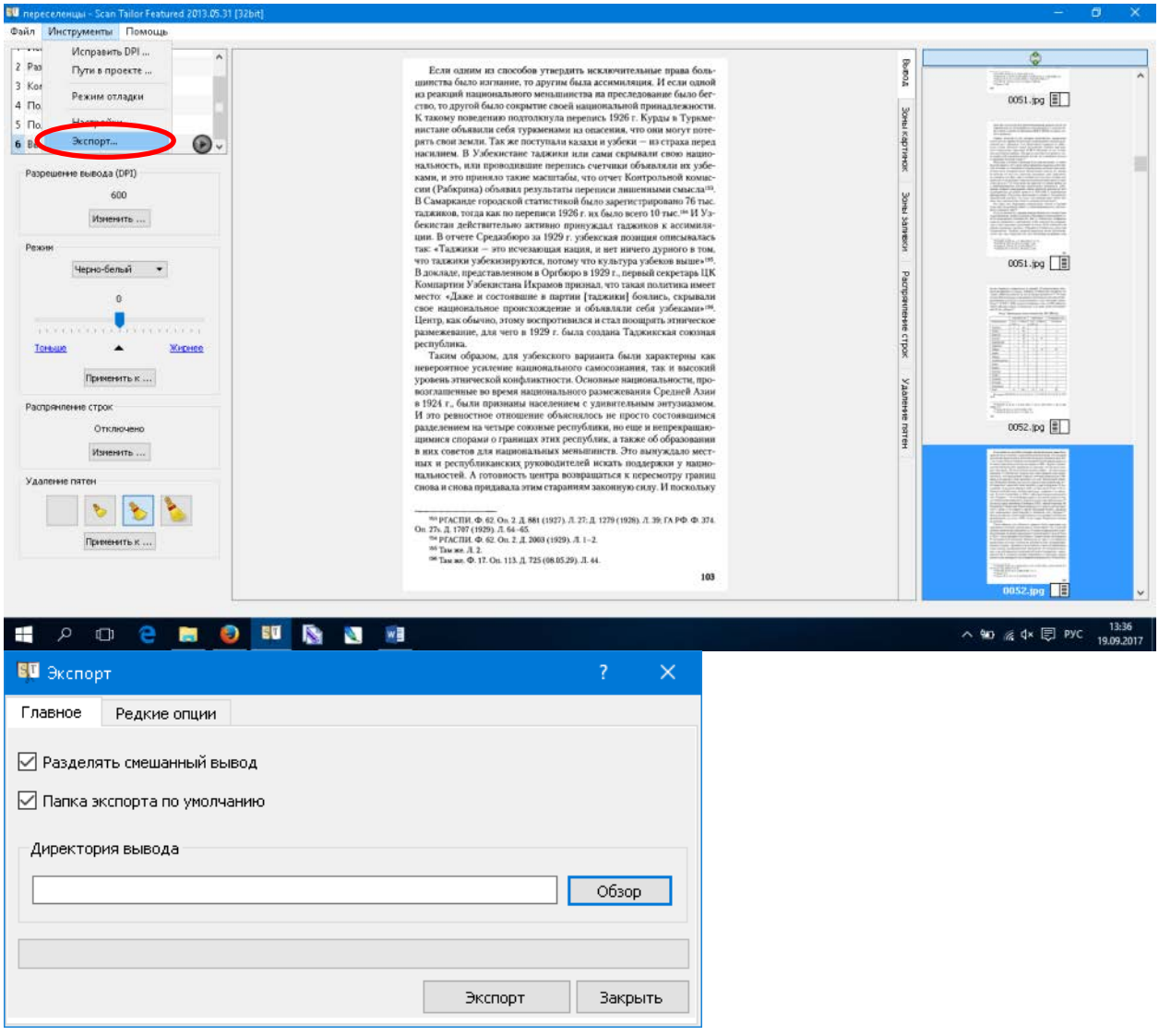


Рисунок 29. Экспорт разделенных сканов в 2 подпапки. Оставляем все параметры по умолчанию и жмём «Экспорт».

После экспорта разделенных сканов страница книги с рисунка выше будет выглядеть так:

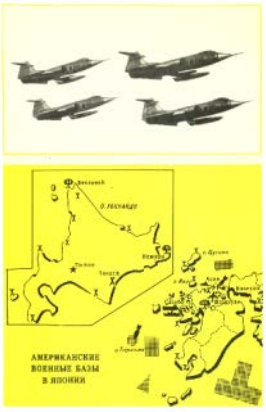



Рисунок 30. Первая страница в папке «1» (черно-белая). Первая страница в папке «2» (цветная).

После экспорта разделенные сканы можно будет кодировать разными алгоритмами, а затем объединить в готовую книгу.

Сложные случаи

Чем красивее книга, тем сложнее её обрабатывать. После бинаризации текст становится абсолютно черным, и если в книге есть цветные буквы, то их приходится выделять как картинки (я знаю, что есть способы раскрасить буквы в DJVU, но я их не освоил). Поскольку вместе с текстом при ручном выделении неизбежно вы захватите ещё немного фона, который может не быть абсолютно белым, после вывода картинок в папку 2 этот фон можно будет удалить в фотошопе с помощью инструмента «волшебная палочка» или «цветовой диапазон».

Детские и подарочные книги имеют цветной фон страниц. Возможно, есть хороший алгоритм разделения текста и фона, но мне он неизвестен. Если мне попадается такая книга, приходится выделять всю страницу как одну большую картинку.

Иногда картинки из папки 2 перед кодирование нужно еще допиливать в фотошопе (например, увеличивать контрастность или делать тоновую коррекцию).

Опыт показывает, что выправить в программах обработки изображений можно всё, кроме размытости. Размазанные буквы не лечатся. Можно увеличить резкость, но, как правило, после бинаризации результат всё равно удручает. Поэтому если вы фотографировали книгу в потёмках трясущимися руками на смартфон, то просто не делайте бинаризацию. Выбелите фон, прогоните сканы через scan tailor, сделайте экспорт в формате «цветной/серый» либо «смешанный» и соберите из полученных tiff'ов электронную книгу в djvu без сжатия, т.е. используя профиль кодирования «photo 600». Получите здоровенный файл, но его хотя бы можно будет читать. Как кодировать в djvu, и где там профиль photo – об этом в следующем разделе.

Кодирование в DJVU

Вопрос о том, какой формат электронных книг – DJVU или PDF – лучше, является почти философским. PDF гораздо более распространен, а программы для работы с ним разрабатываются серьезными конторами. Программы для работы с DJVU обычно созданы энтузиастами на коленке, многие люди вообще не в курсе, что такой формат существует, некоторые электронные ридеры до сих пор не умеют его читать. Но при этом DJVU был создан специально для сканированных изображений, файлы получаются меньше, открываются быстрее, а качество в DJVU получается ничуть не хуже. Лично я так и не научился делать PDF приемлемого качества и приемлемого

размера, поэтому определенно предпочитаю DJVU. Несколько сотен мегабайт для книги – это всё-таки перебор. Конечно, для создания PDF из подготовленных сканов нужна одна программа, а для создания DJVU – три, но после второй созданной книги это перестанет вас пугать. Тем более, что все три программы работают очень быстро. PDF я использую, только если меня об этом специально просит владелец книги.

Для кодирования в DJVU вам нужны три программы: **DJVU small, DJVU imager** и **FR11 DjVu Text Layer Crutch** (все три – бесплатные), а также **Abbyy Finereader версии 11 и выше** для распознавания текста и **PDF&DJVU Bookmarker** для создания активного оглавления.

Программой **DJVU small** мы создаем основу будущей книги – кодируем текст. Программе надо указать путь к папке «1», созданной во время экспорта сканов из skan tailor'a, и задать профиль кодирования «bitonal 600 dpi». Я пользуюсь программой уже несколько лет, и так до конца не освоил, зачем там остальные профили (не считая профиля photo, который позволяет создавать djvu книги вообще без сжатия). Если программа запускается первый раз, также понадобится указать путь для сохранения результата (выходную папку). В результате работы программы получается файл djvu.encoded.djvu.

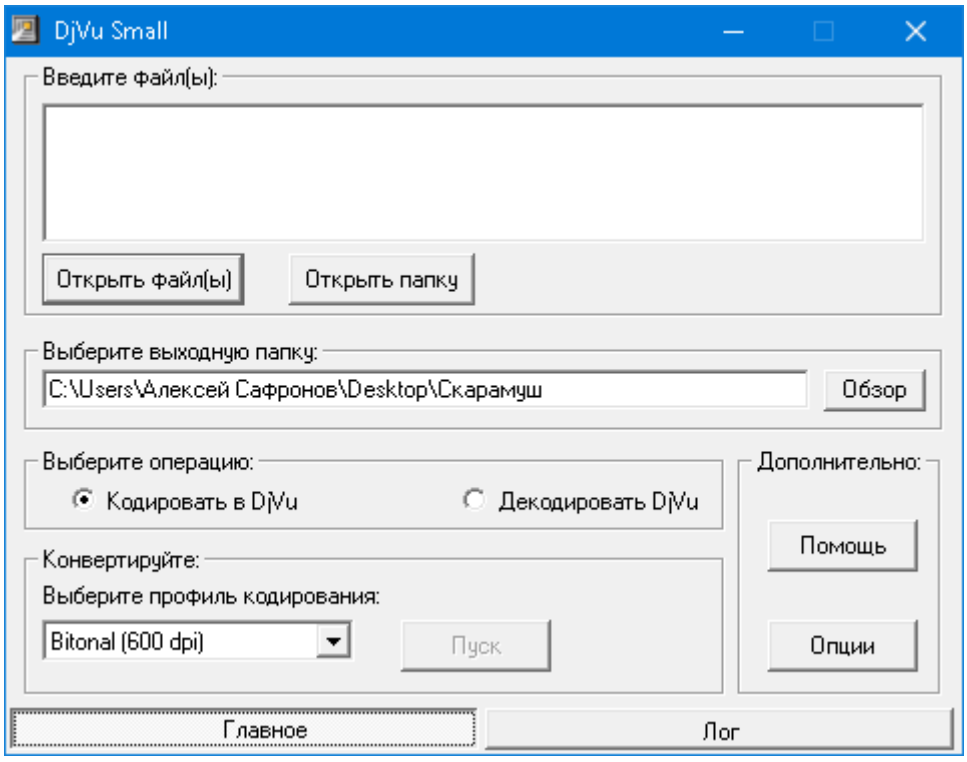


Рисунок 31. Настройки программы DjVu Small для кодирования файлов из папки «1».

При необходимости той же программой DJVU small можно разбирать (декодировать) djvu-книги на отдельные страницы (файлы). Для этого надо поставить галочку напротив «декодировать DjVu» и при необходимости в «опциях» настроить формат вывода (я декодирую в tiff) и диапазон страниц.

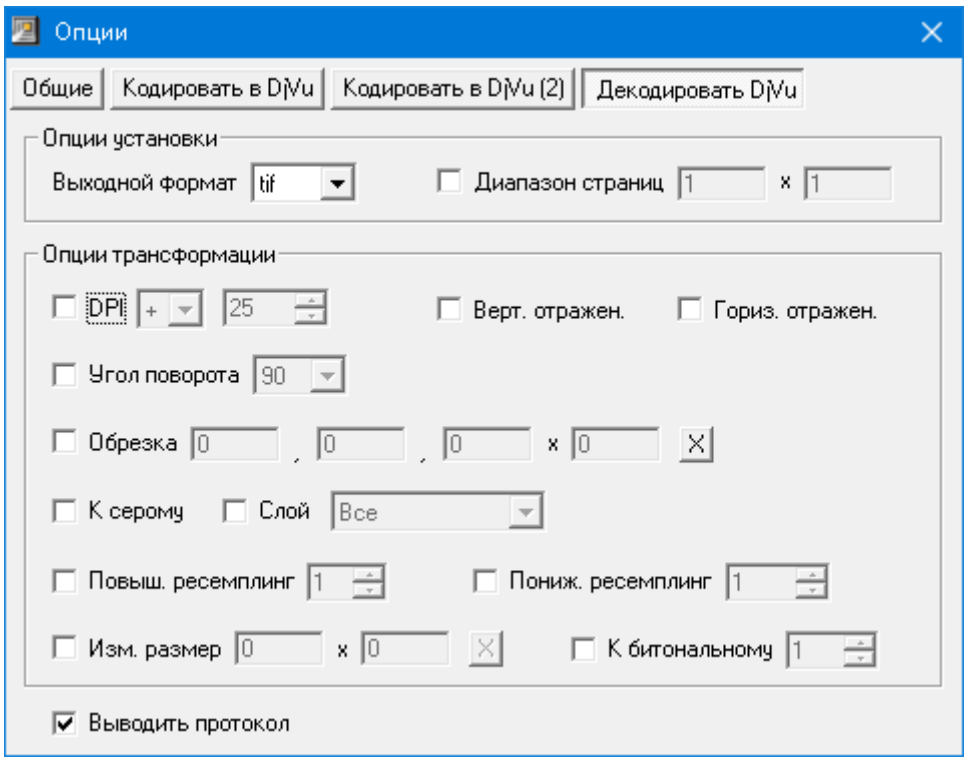


Рисунок 31. Вкладка «Опции» программы DjVu Small с настройками для декодирования djvu-книг

Теперь в нашу книгу надо вставить картинки. Как минимум фото обложек. Для этого используется программа **DJVU imager.** Ей, соответственно, надо указать путь к папке «2». Настройки по умолчанию в большинстве случаев трогать не нужно. Если программа отказывается открывать файлы, поставьте в «опциях» галочку «Произвольные файлы». Сначала жмем кнопку пуск, после чего программа кодирует все картинки и показывает их нам. Теперь надо прописать путь к расположению книги, в которую будут вставляться картинки. Полный путь достаточно указать один раз. После кодирования становится активной кнопка «вставить в DjVu». Жмём её. После вставки получаем файл djvu.encoded.out, который фактически является уже готовой книгой.

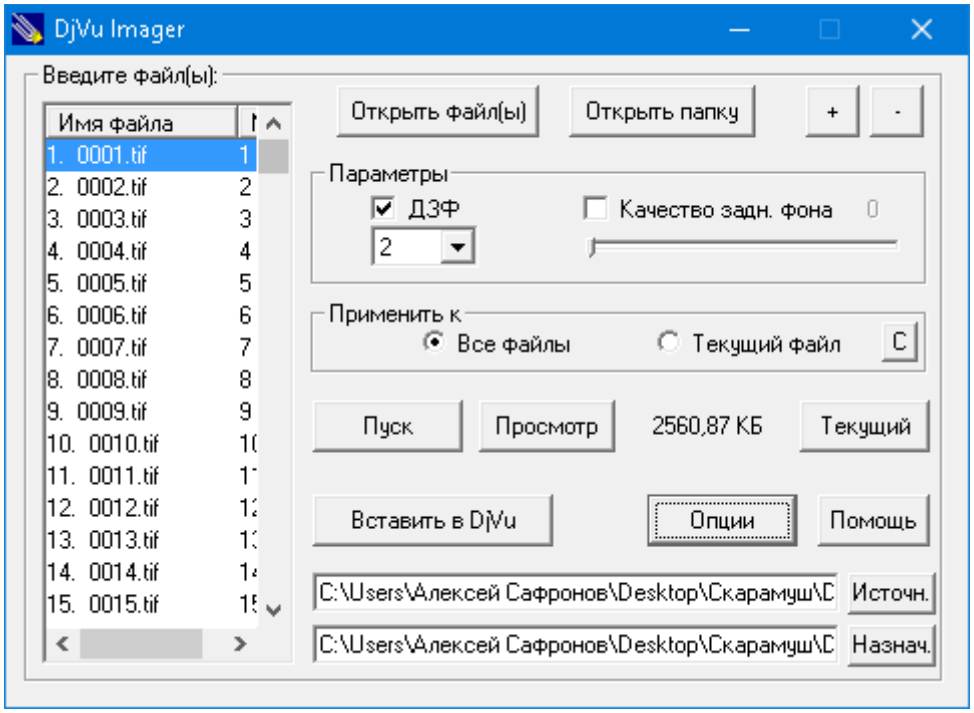


Рисунок 32. Настройки программы DjVu Imager для кодирования файлов из папки «2».

Теперь в неё надо добавить OCR-слой (слой текста, позволяющий искать по книге поиском и копировать цитаты оттуда) и активное оглавление.

Файнридер версии 11 и выше умеет открывать djvu-файлы и распознавать текст в них, но к сожалению он сохраняет распознанный текст не в тот же файл, а создает новый djvu-файл с текстовым слоем. При этом встроенный кодировщик файнридера оставляет желать лучшего, поэтому новый djvu-файл может оказаться не слишком хорошего качества.

Для борьбы с этой напастью умельцами была разработана программа FR11 DjVu Text Layer Crutch которая умеет переносить текстовый слой из одного djvu-файла в другой. Программу можно скачать здесь: https://yadi.sk/d/IAuQbT2c3QfdnE

Таким образом, нам надо сначала распознать текст файнридером, затем сохранить результат распознавания в новый djvu-файл, потом скопировать текстовый слой из нового файла в старый, полученный на предыдущем этапе, а потом удалить новый djvu-файл за ненадобностью.

Запускаем файнридер, указываем ему путь к папке «out» нашего проекта scan tailor. Ждем, пока все страницы добавятся в проект. Выделяем все страницы (ctrl+A), жмём «распознать». Ждем, пока закончится распознавание, потом сохраняем результат в djvu нажатием клавиши «Сохранить» (рисунок 33). Первый раз может потребоваться выбрать формат сохранения, для этого рядом с кнопкой «сохранить» есть стрелочка вниз, при нажатии которой открывается выбор форматов.

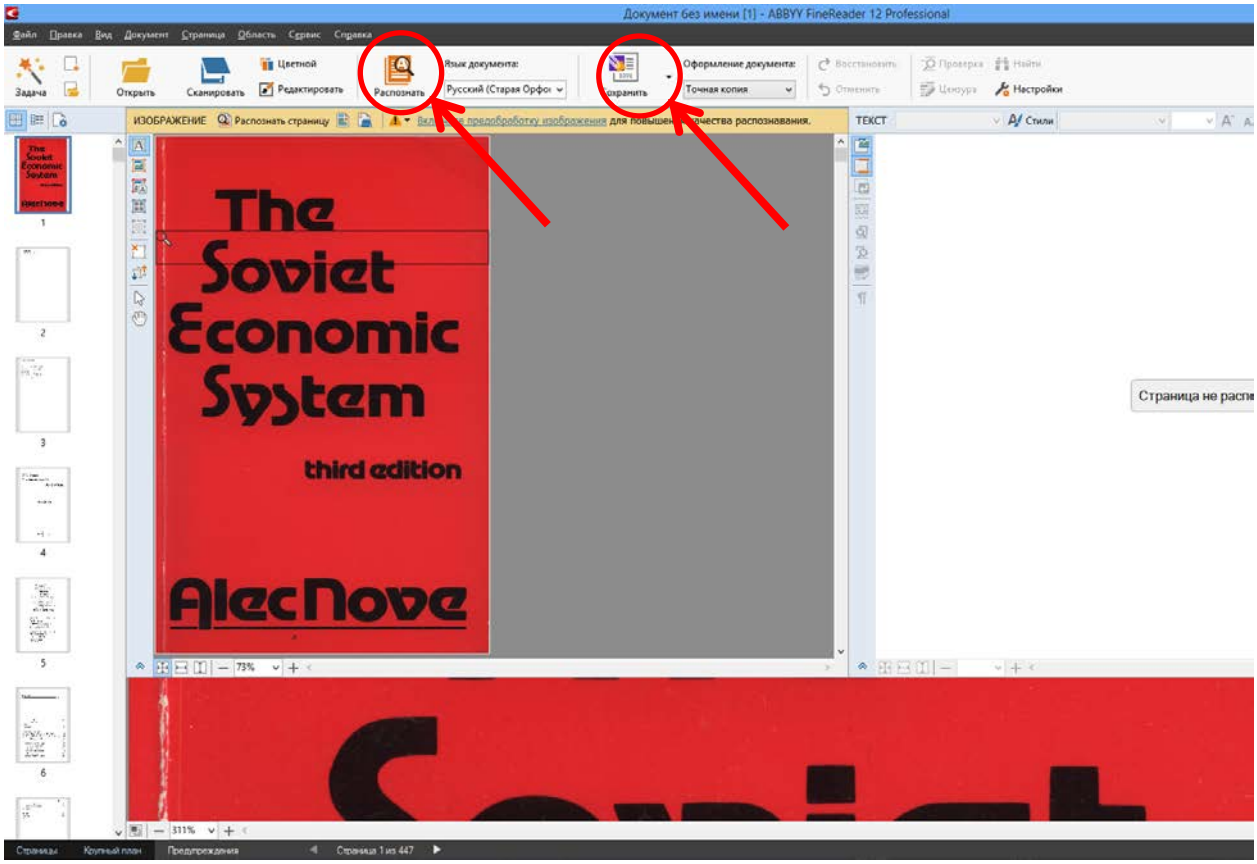


Рисунок 33. Распознавание текста в программе Fine Reader 12 и сохранение результата в Djvu.

Теперь нам надо открыть программу FR11 DjVu Text Layer Crutch и указать ей два djvu-файла: файл, созданный файнридером (с текстовым слоем) и файл, полученный путем использования программ djvu small и djvu imager (рисунок 34).

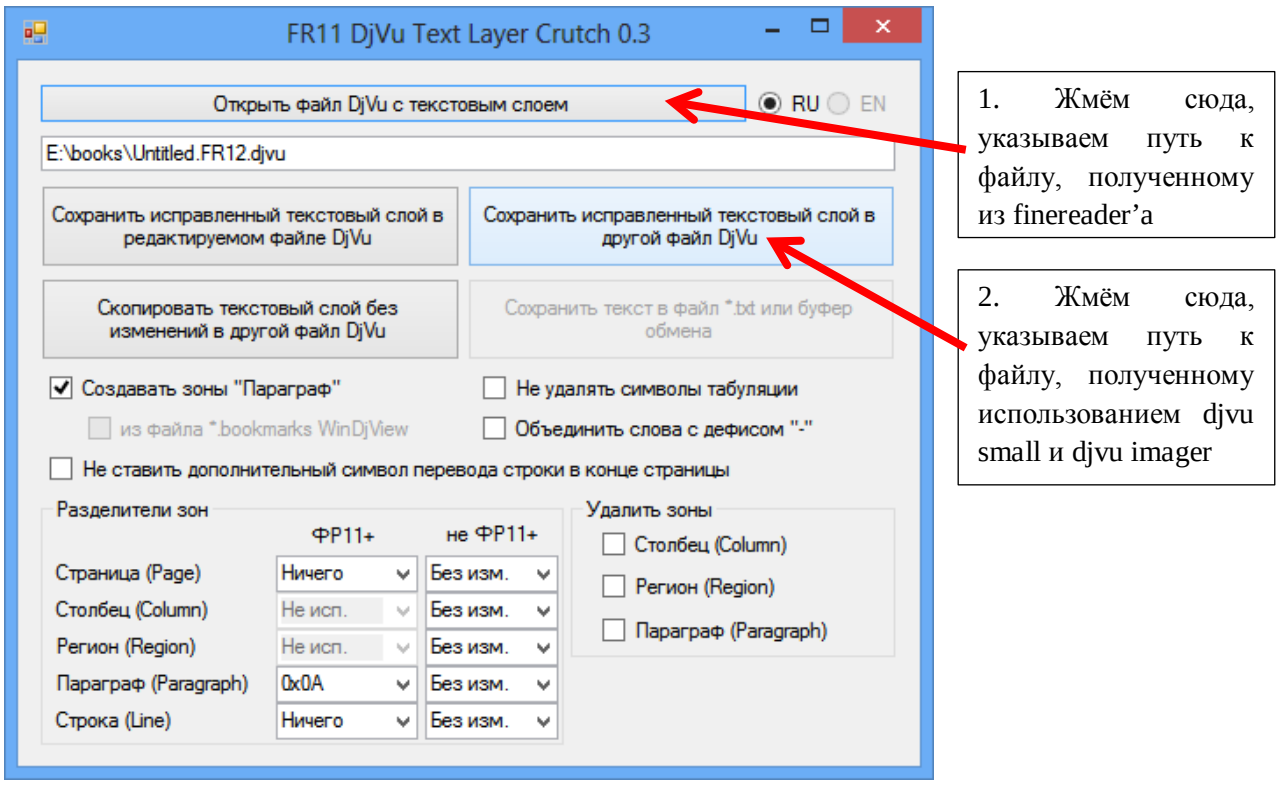


Рисунок 34. Окно программы FR11 DjVu Text Layer Crutch

Теперь в наш файл добавлен текстовый слой. Главное потом не перепутать, какой вариант кодировался файнридером (с посредственным качеством), а какой мы получили использованием djvu small и djvu imager.

Завершающим этапом является вставка в книгу активного оглавления, которое при чтении книги видно сбоку программы и позволяет по щелчку переходить к нужной главе. Очень полезная вещь, особенно в толстых книгах, а также в сборниках статей.

Прекрасной программой для создания оглавления является **PDF&DJVU Bookmarker**. Программа скачивается и устанавливается вместе с подробнейшей инструкцией по использованию, к которой я и отсылаю читателя. В неё удобно копировать оглавление, распознанное файнридером на соответствующих страницах книги на предыдущем этапе, поэтому после сохранения проекта файнридера не спешите закрывать его. Кроме того, в программе есть опция копирования оглавления из одной книги в другую. Если вы создаете сразу два файла – DJVU и PDF, то она оказывается весьма полезной.

Не забудьте переименовать книгу. Создастся она у вас по умолчанию с именем «djvu encoded.out» или каким-то наподобие. Я свои книги обычно называю в формате «Фамилия И.О. – Название год издания». Если вы забудете переименовать книгу, то при создании следующей книги программы перезапишут файл «djvu encoded», и прежняя работа будет потеряна.

Ваша DJVU-книга готова. Не забудьте выложить её в интернет.

Кодирование в PDF

С PDF всё проще. Если у вас уже установлена программа **Adobe PDF creator**, то вы можете зайти в папку «out» проекта Scan tailor (это в этой папке при необходимости создается подпапка export с подпапками 1 и 2), выделить все файлы (ctrl+A), и по правому щелчку мыши будет доступна

опция «объединить файлы в PDF». Выбираем её, соглашаемся со всем, что предлагают дальше, и получаем готовый PDF. В программе **Adobe PDF creator** есть свой модуль распознавания текста, поэтому, когда pdf будет готов, найдите справа в «инструментах» распознавания текста, согласитесь там со всем, и получите PDF с текстовым слоем.

Не забудьте сохранить полученный результат, так как поначалу вам покажут временный файл во временной системной папке **PDF creator**'а.

Обычно из одних и тех же сканов файл pdf получается в 20 раз больший по размеру, чем djvu. В **Adobe PDF creator** есть опция «сохранить как файл меньшего размера», но от неё страдает качество.

PDF creator позволяет и оглавление в книге создавать без дополнительных программ, но **PDF&DJVU Bookmarker** настолько хорош и удобен, что я всегда пользуюсь им.

Ваша PDF-книга готова. Не забудьте выложить её в интернет.

16.12.2017

Vas_s_al