

Создание электронных книг в формате DjVu

Введение	1
1. Сканирование	1
2. Обработка	3
3. Обработка фото и цветных иллюстраций	9
4. Кодирование	10
5. Создание текстового слоя	12
6. Добавление обложки	13
7. Оглавление	13
7. Используемые программы и где их взять	15
Заключение	15

Введение

Эта мини инструкция в картинках, описывающая полный цикл создания электронной версии научно-технической книги, и предназначена для человека, искренне захотевшего сделать приемлемого качества е-книгу, но не знающего с чего начать. Важно понимать, что существует немало апробированных методов создания достаточно качественных е-книг, все они характеризуются тем, что на выходе книга, как правило научно-техническая, имеет разрешения 600 dpi ч/б (все книги в 300 dpi ч/б, несмотря на все старания создателей, явно проигрывают в качестве).

Рассматриваемая здесь метода, основана на сканировании в **300 dpi, в градациях серого** (600 dpi ч/б будет после обработки). По этому поводу следует заметить, что уменьшение геометрического размера сырого скана в 4 раза, по сравнению со сканированием в 600 dpi, практически компенсируется увеличением глубины цвета в 8 раз (зато скорость сканирования возрастает в 2 раза ☺), а также уменьшением количества паразитного мусора (чистить практически не надо будет).

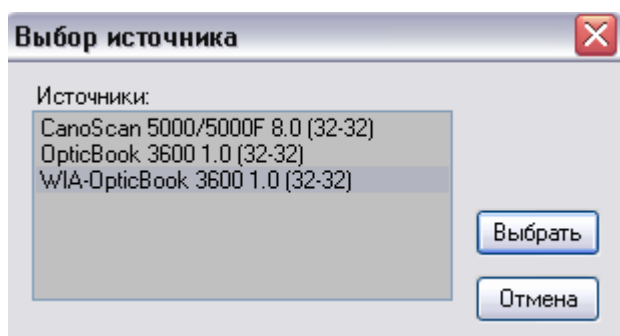
1. Сканирование

Беря в руки увесистую книгу, многие думают, что отсканировать ее может только маньяк. Совершенно верно, именно так. Без применения научно-организованного подхода, любая работа превращается в мучение, но, сделав работу незаметной, хоть большого удовольствия и не получишь, но дело сделаешь.

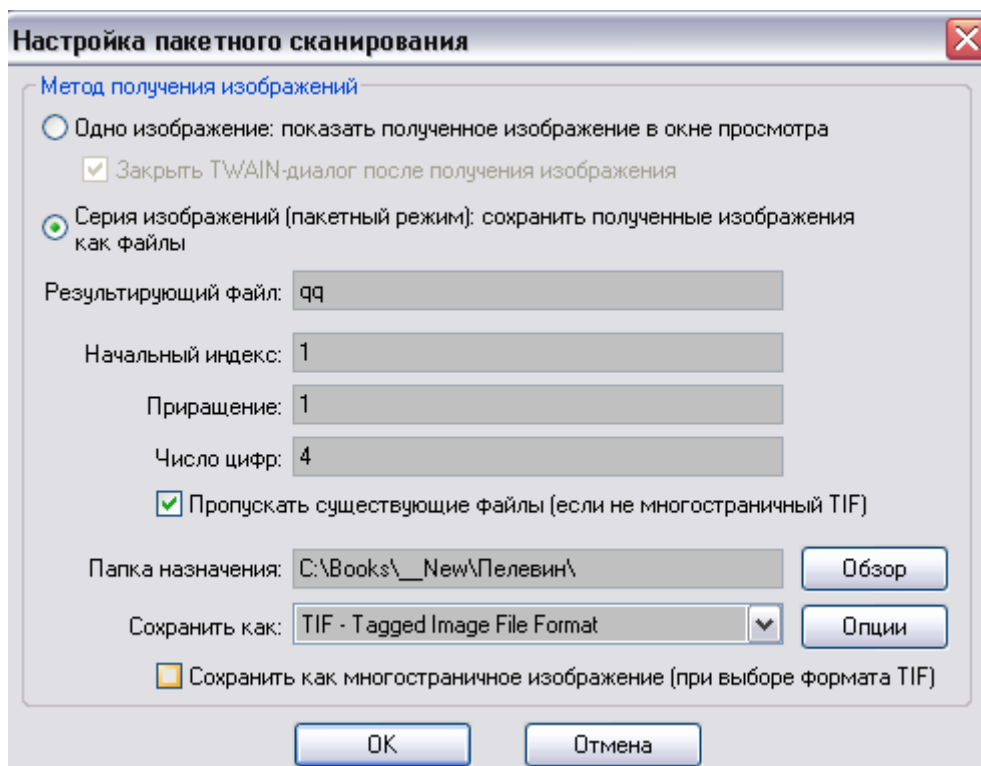
Для сканирования сгодится любая программа, способная взаимодействовать с TWAIN драйвером сканера и сохранять отсканированные изображения на диск, нумеруя их удобным способом. Сойдет любой просмотрщик графических файлов: **ACDsee, IrfanView, XnView...** Если ваш сканер поддерживается программой сканирования **VueScan**, можете использовать и ее.

Например, в **IrfanView** (скачайте свежую версию этой бесплатной программы) это выглядит примерно так:

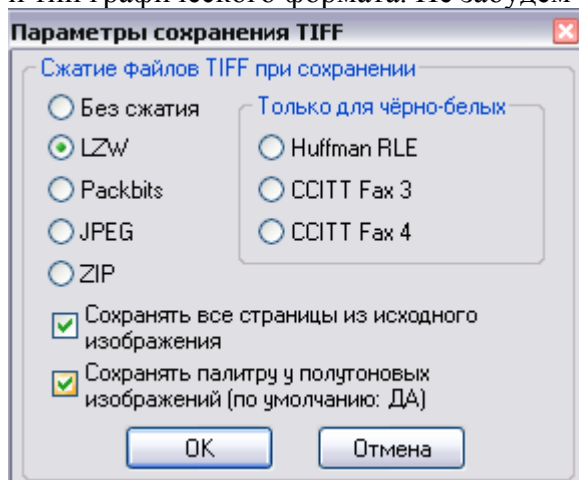
- В меню **Файл** жмем пункт **Выбрать TWAIN-источник...**



- Далее, там же, выбираем пункт меню **Получить изображение/пакетное сканирование...**



здесь выбираем как будут нумероваться файлы сканов, где они будут складироваться и тип графического формата. Не забудем проверить **Опции** графического формата:



можно выбрать или **Без сжатия** или **LZW** (внимание, не все программы корректно с ним работают), в последнем случае размер файла на выходе будет примерно в два раза меньше. Можно, наверное, и **ZIP**, но это проверьте самостоятельно.

- жмем на кнопку ОК и переходим в окно TWAIN Вашего сканера

Сама техника сканирования незатейлива:

- берется книга, кладется разворотом (т.е. двумя страницами) на стекло, прижимается если надо сверху рукой (это быстрее, чем использовать груз).
- делается предварительное сканирование
- картинка, если это возможно, в окне сканирования, разворачивается на 90 градусов (в нормальное положение)
- выбирается область сканирования с некоторым запасом, как правило по горизонтали (по вертикали трудно промахнуться)
- мышкой жмется кнопка основного сканирования
- после того, как данный разворот отсканирован, во время обратного движения каретки сканера, переворачиваем страницу книги, кладем на то же место и ждем

опять на левую кнопку мыши (курсор ведь остался на кнопке сканирования), и так пока книга не кончится.

Т.е. идея проста, сканируем развороты в слепую. Этим достигаем максимальной скорости сканирования, которая ограничена только техническими характеристиками сканера, и полной свободой головы. Таким образом, во время сканирования, Вы можете заниматься многими другими вещами, да хоть кино посмотреть.

Небольших перекосов, отсканированных страниц, бояться не стоит, это будет исправлено при последующей обработке, но все же надо соблюдать аккуратность. Желательно все же таки серединку прижимать посильнее, исправление геометрических искажений строк здесь не будет рассмотрено.

Не забываем, что сканируем с разрешением **300 dpi и в градациях серого (gray scale)**, если будете сканировать в черно-белом режиме при 300 dpi, то просто потеряете время (хорошая книжка уже не получится).

На выходе этого этапа получаем так называемый сырой материал – файлы в формате **tiff с разрешением 300 dpi в градациях серого**, обычно размер каждого файла, без использования сжатия, составляет примерно 8 мегабайт (4 при LZW).

Скорость сканирования может достигать до 200 и даже более разворотов (400 страниц) в час, на сканере со скоростью 16 секунд на сканирование A4, т.е. сканирование среднестатистической книги, займет не более 2 часов времени! Ну а если у Вас Plustek OpticBook 3600, то за час можно отсканировать более 500 страниц (250 разворотов).

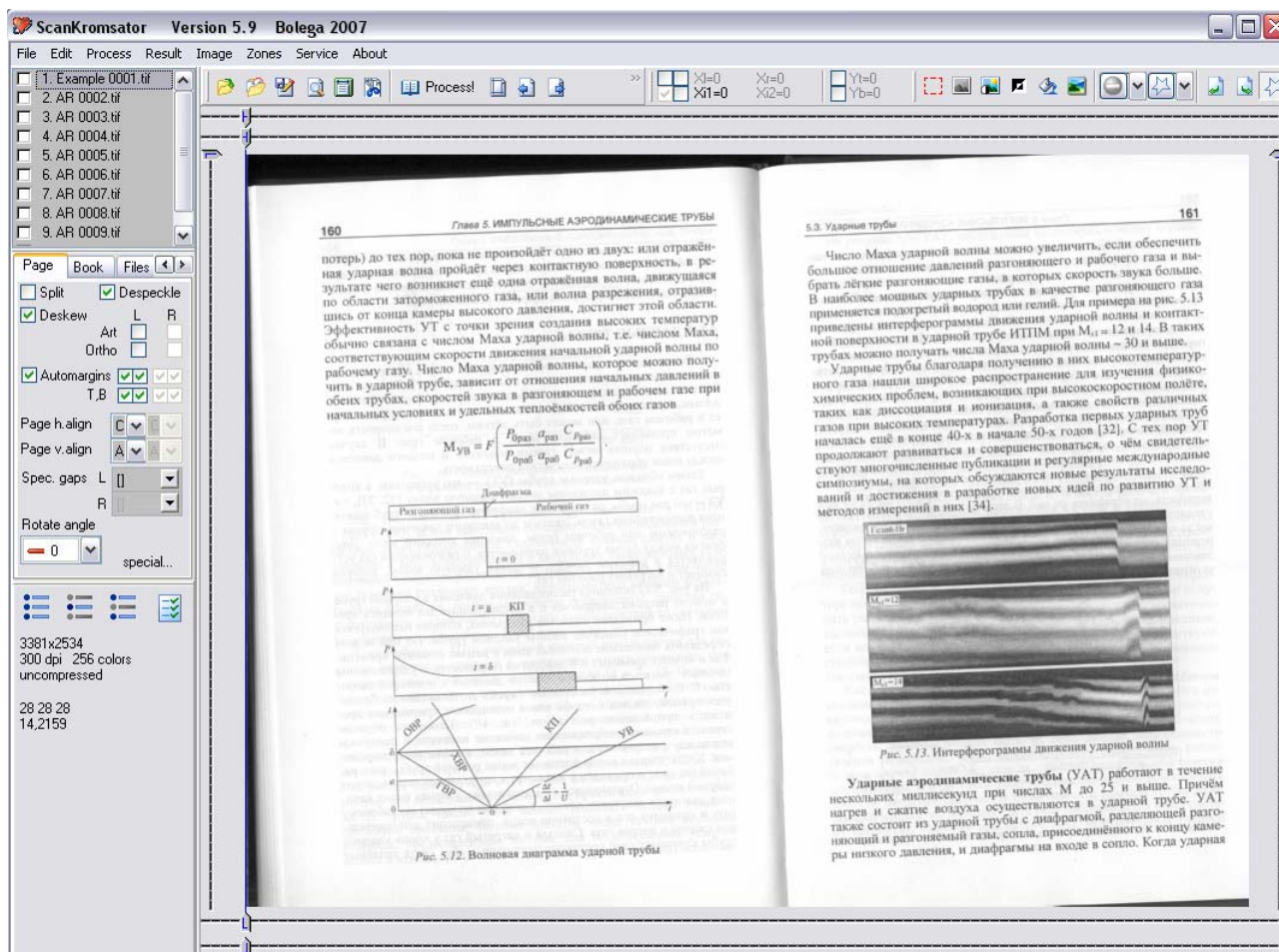
Обычно первый и последний разворот книги содержат по одной странице. Ну, так и сканируем их по одной, т.е. все-таки придется сделать 3 предварительных сканирования на книгу ☺.

2. Обработка

За обработку сырого материала отвечает замечательная, притом совершенно бесплатная, программа **ScanKromsator** от **bolega** (тут убедительная просьба, не надо сразу же бросаться и писать ему письма о том, как улучшить, углубить, да и просто спасибо, наверное, то же не надо посылать, просто сделайте **хорошо** несколько книг и поделитесь ими).

ScanKromsator это мощный инструмент, предназначенный для обработки сканированного материала, с целью создания качественных е-книг, со многими полезными и не очевидными для новичка функциями. Поэтому, просто следуйте пошаговой инструкции и все получится.

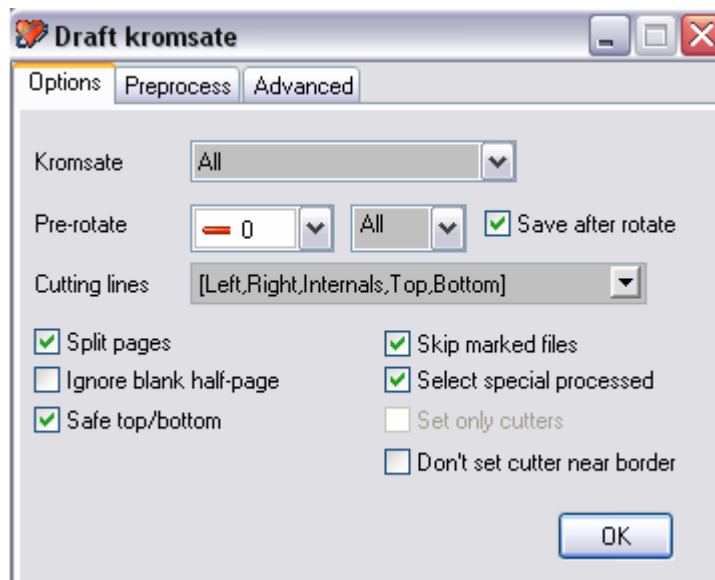
а) Запускаем программу и загружаем в нее файлы (список файлов слева сверху, под этим списком панель инструментов):



б) Выбираем путь для вывода результатов (закладка **Files**), по умолчанию out текущего каталога, тут же можно назначить способ нумерации выходных файлов. **очень важно, назначить выходное разрешение 600 dpi.**

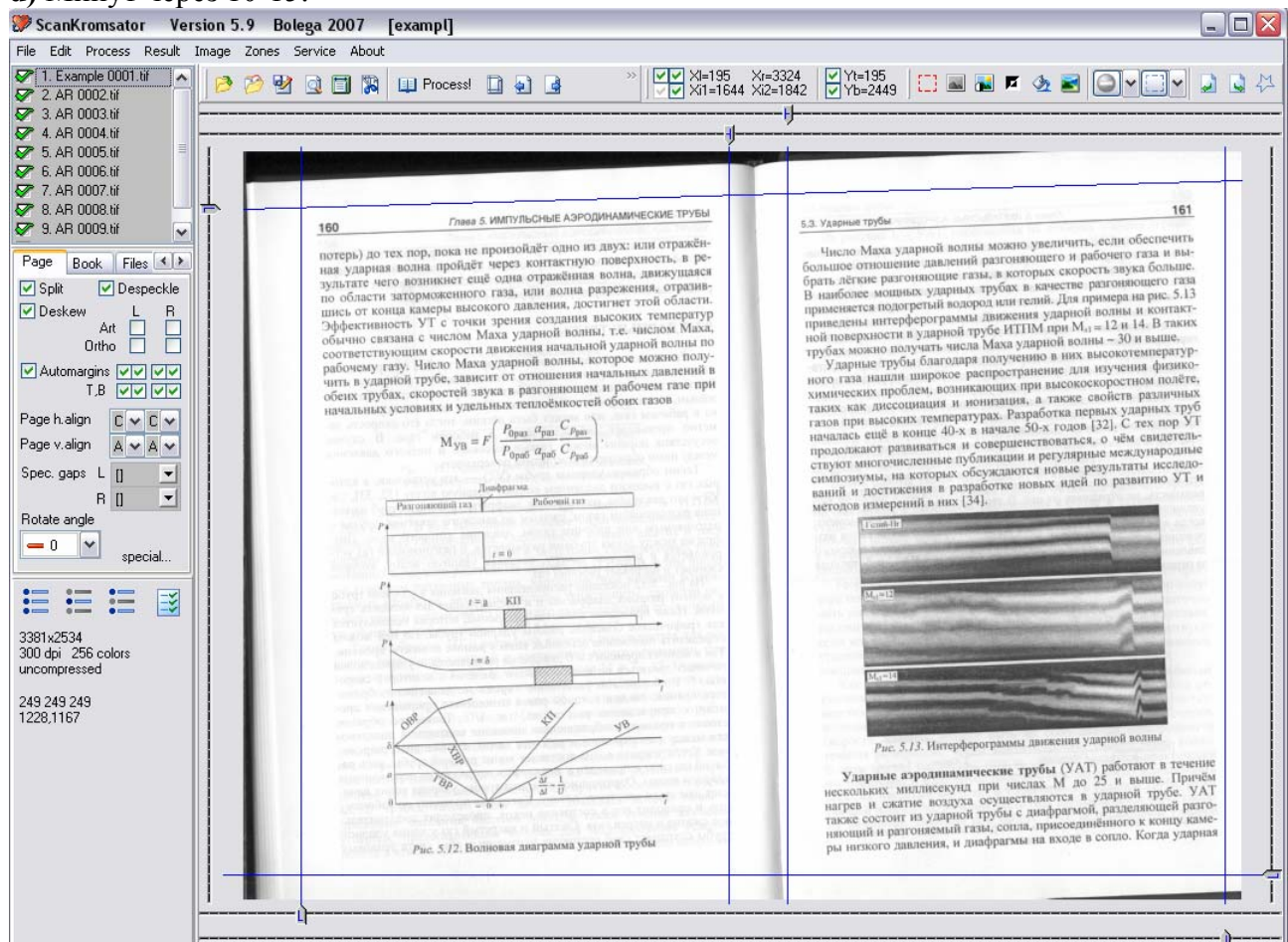


в) Приступаем к черновому «кромсанию»: Находим левее кнопки с надписью **Process**, кнопку с ножницами (**Draft kromsate**), нажимаем, появляется окно диалога

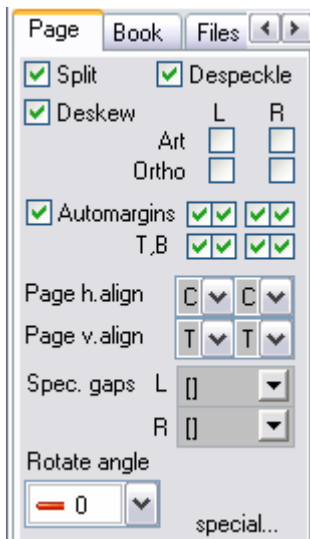


ставим галочки на **Split pages** и **safe top/bottom** и ждем кнопку **OK**.
 (если первая и/или последняя страницы одинарные, т.е. не разворот, то можно предварительно покромсать их отдельно (поле **Kromsate** = **Current**), соответственно не надо для них ставить галочку **Split pages**)

d) Минут через 10-15:



Обратите внимание на синенькие полосочки, это резак (по которым Вы безошибочно отличите это программу от других ☺), за их пределами все будет безжалостно отрезано, а данная страница будет разделена на две (см. центральные резак). Посмотрите на то, что рядом с названиями страниц появились **зеленые галочки!**



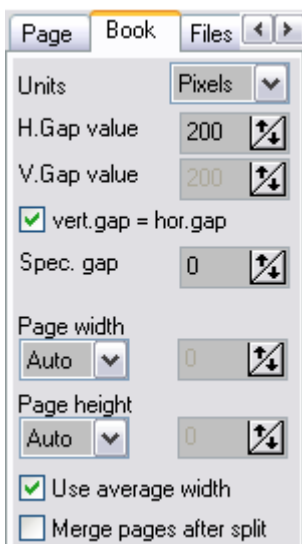
е) Это короткий, но очень важный этап – расстановка опций. Для этого пройдемся по закладочкам (слева в окне программы).

Pages. На ней выставляем способ центрирования. По умолчанию стоит **A** – автомат, это значит поместить изображение в верхний левый угол. Но, как правило (это у меня так) горизонтальное выравнивание ставится по центру (**Page h.align**) **C**, вертикальное в низ (**Page v.align**) **B** или вверх **T** это зависит от оригинального форматирования книги.

Despeckle – убирание мелкого мусора.

Deskew – выравнивание наклона страницы, если в результате кромсания страница получится криво выровненной, то ее можно переделать с помощью метода **Art** (включение этого метода для всех страниц замедляет процесс) или **Ortho** если текст на данной странице развернут на 90 градусов.

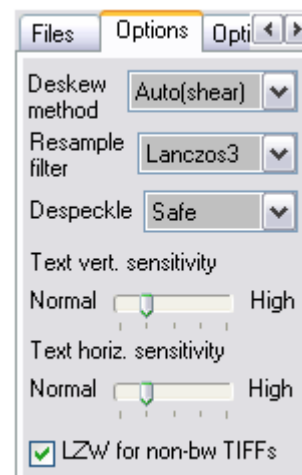
Чтобы опция была применена ко всем страницам, при выборе ее удерживаем **Ctrl**. Аналогично действуйте при выборе остальных опций, которые применяются ко всем страницам сразу.



На закладке **Book** выставляем размеры выходных страниц, оставляем **Page width** и **height** в **Auto**. В поле **H.Gap value** ставим 200 (или 250) pixels, это значение обычно для обработки в 600 dpi, но если Вам хочется других размеров полей, то можете подобрать это значение по своему вкусу.

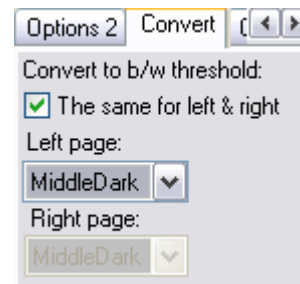
В закладке **Files**, как было сказано выше, ставим выходное dpi **600** (иначе ничего хорошего не получится). Это архи важно, от этого зависит весь окончательный результат.

Во вкладке **Options**, ставим **Deskew method** = **Auto(shear)**, для **Despeckle** метод **Safe** или **Fine+Normal** это интеллектуальный метод очистки. Например, он не вычищает точки над **i** и **j**. Также можно подвинуть ползунки для **Text sensitivity** на два три деления, чтобы резаки не обрезали отдельно стоящие номера страниц.

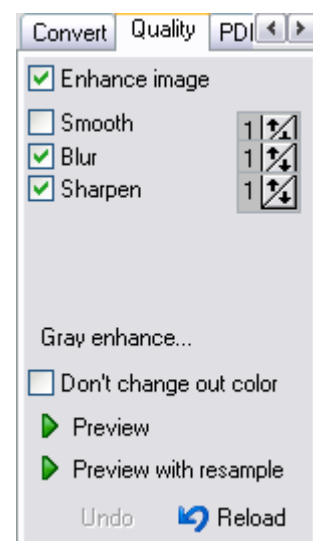


Options 2 пропускаем.

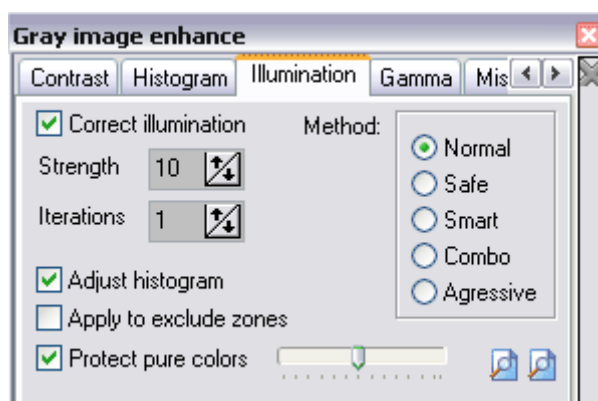
Вкладка **Convert** – выставляем порог для преобразования из градаций серого в черно-белый. Для **Convert to b/w threshold** выбираем **MiddleDark**. Не забываем удерживать Ctrl при выборе опции, предназначенной для всех страниц. Но никто не мешает провести эксперимент для своего скана и выбрать другой параметр.



Ну, наконец, последняя, но очень важная вкладка **Quality**. В **Enhance image** ставим галочки для **Blur** и **Sharpen**, значения для них обычно 1 или 2 (набор этих опций и их значения не догма, можете поэкспериментировать, но для начала поставьте как на рисунке), для 2 результат будет пожирнее, выбирайте исходя из шрифта, сканируемой книги.

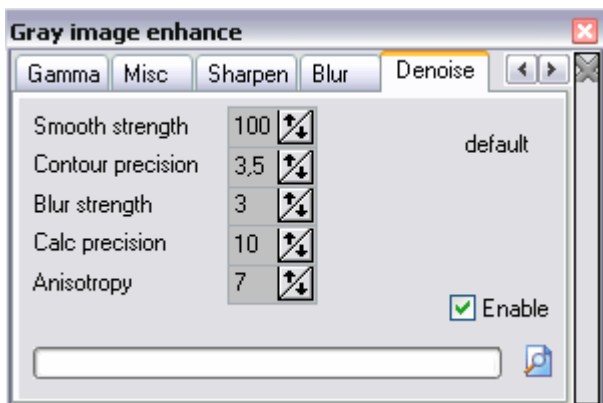


И опять очень важно, если у вас сканы в градациях серого, то ждем на **Gray enhance** и появляется диалог **Gray image enhance**, переходим на вкладку **Illumination** где ставим зеленую галочку на **Correct illumination**.



По этой опции происходит выравнивание освещенности (особенно важно это для центра разворота), что убирает черные полосы и кучу мусора. Незаменимая штука.

Здесь же листаем дальше и находим вкладку **Denoise** ставим галку на **Enable**, а параметры как на рисунке:

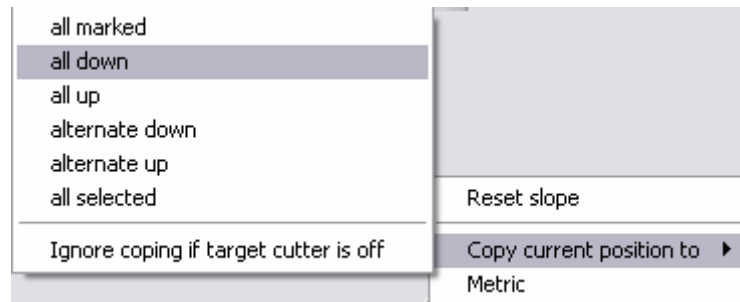


Все, все параметры для кромсания выставлены, чтобы каждый раз не мучаться, можно создать свой профиль **File->Options...**

f) Самый скучный, но к счастью не очень долгий этап. Надо пройтись по всем страницам, с целью проверки правильности расстановки резаков и выделения иллюстраций, если они имеются.

Если Вы увидите, что для какой либо страницы резаки установлены не правильно, то их надо поправить. Передвигаем резаки, если надо меняем способ центрирования для данной страницы (если текст на странице развернут на 90°, то для данной странице ставим **Deskew = Ortho** на закладке **Pages**).

Оптимально это делается так: левая рука отвечает за листание – кнопки **q** и **w**, правая за мышью, которой мы передвигаем, если надо резаки. Если Вы уверены, что для части страниц положение резака будет одинаково, то Вы можете скопировать их положение, нажав правую кнопку мыши на резаке, выберите нужную опцию (**Copy current position to**).



Бывает, что страница расположена под углом, или тень на развороте расширяется, для таких случаев можно устанавливать косые резаки, просто, удерживая шифт, передвигаем резак за его кончик, это быстрее, чем в последствии в ручную чистить страницы. Для наглядности, в качестве примера, была выбрана криво сосканированная страница, верхний резак как раз «кривой».

g) Если на странице есть ч/б фото, цветная или полутоновая иллюстрация, то в кромсаторе предусмотрен специальный режим их обработки, так называемые **Picture Zones**. Выделяем мышью иллюстрацию прямоугольником и просто нажимаем на кнопочку **Mark as picture zone**, если иллюстрация имеет неправильную форму или прямоугольную, но косо отсканированную (как в примере), то можно использовать **Polygon selection**, иконка в виде кривой звездочки.



После кромсания, таким образом выделенные картинки будут помещены в отдельные файлы, подробнее смотрите дальше.

е) Кстати, знаете ли Вы, чтобы все не делать заново, задание можно сохранить (пункт основного меню **File->Save Task**)



ф) Жмем большую кнопку **Process**. Тут появляется предупреждения, в здравом ли мы уме, что меняем разрешение, но нам уже все равно, мы все уже сделали.

Все, теперь дело за компьютером.

Хорошая новость, в новой версии кромсатора процесс кромсания резко ускорился, на компьютере с процессором E6550, кромсание идет со скоростью 6 разворотов (12 страниц) в минуту.

Через некоторое время, минут через 20-30, в указанной ранее папке, нас ждет результат, просматриваем его внимательно, иногда могут быть несколько неправильно выровненных страниц. Их переделываем отдельно.

Совершенно не обязательно кромсать всю книгу сразу, можно делать это по частям. Просто, в последующих порциях, необходимо выставить **Book ->Page width->Fixed** размер предыдущей части. Для определения правильного размера в кромсаторе, обычно, достаточно взять 10÷15 разворотов (страниц).

Особо дотошные, могут почистить остатки вручную, так называемая тонкая очистка. Лично я это не делаю, за исключением убирания библиотечных штампов и записей на полях. Как правило, и так все замечательно. Кстати, в сканкромсаторе есть мощные средства для очистки сканов, можете воспользоваться.

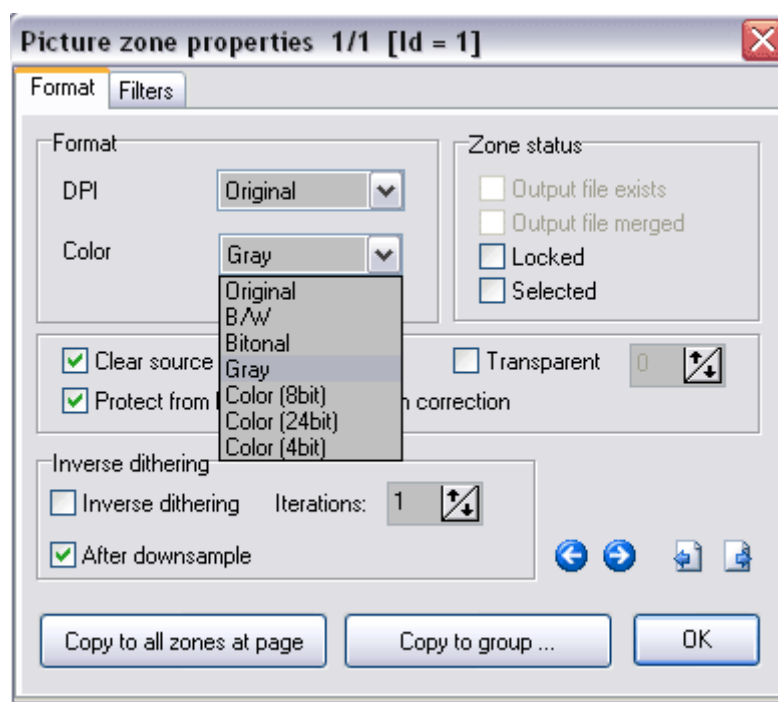
3. Обработка фото и цветных иллюстраций

Описание обработки фото и цветных иллюстраций вынесено в отдельный пункт инструкции, хотя это делается в кромсаторе на этапе проверки правильности расстановки резков. Для обработки таких иллюстраций в последней версии кромсатора были введены так называемые **Picture zones**.

Обводим с помощью мыша иллюстрацию (закключаем ее в прямоугольник) и нажимаем на иконку **Mark as Picture zone** на панели задач.



Настроить параметры **Picture zone** можно «даблкликнув» мышью на выделенном участке, появится диалог настройки **Picture zone properties**, там необходимо выставить цвет (**color**) иллюстрации, по умолчанию стоит **Gray**.



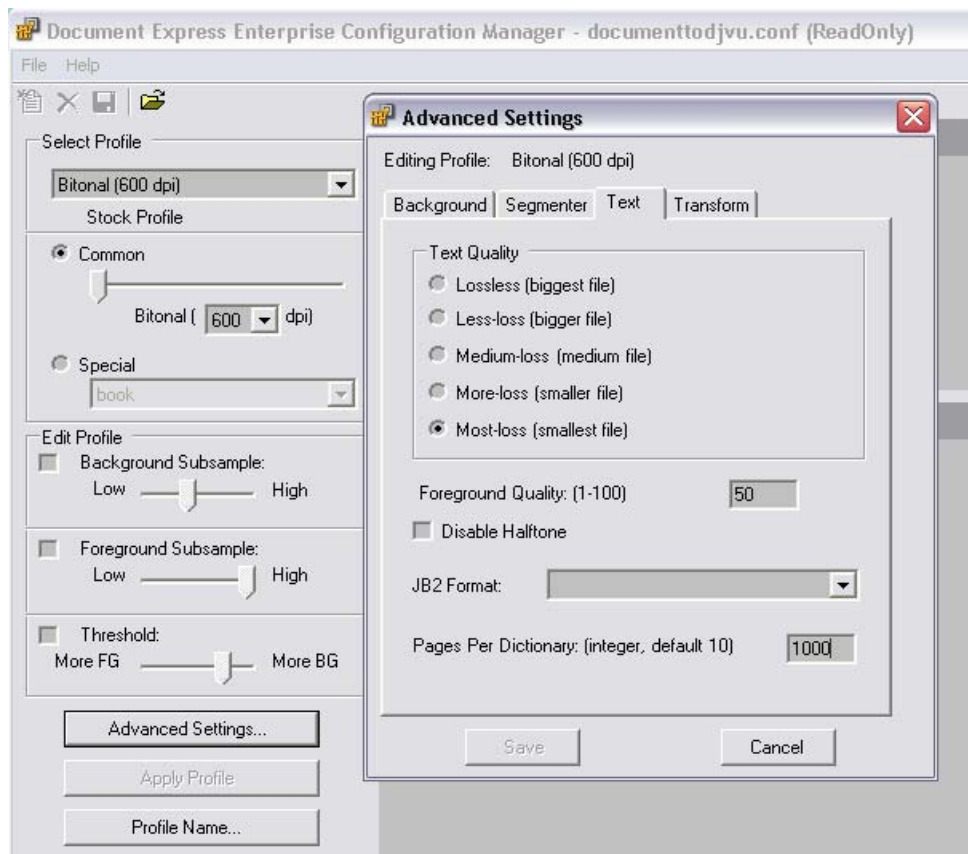
Так как после кромсания иллюстрации выделены в отдельные файлы, это для продвинутых пользователей и эстетов, нам необходимо их объединить со страницами книги. Для этого просто выбираем пункт меню **Zones->Picture Zone->Merge zones...** жмем и готово.

4. Кодирование

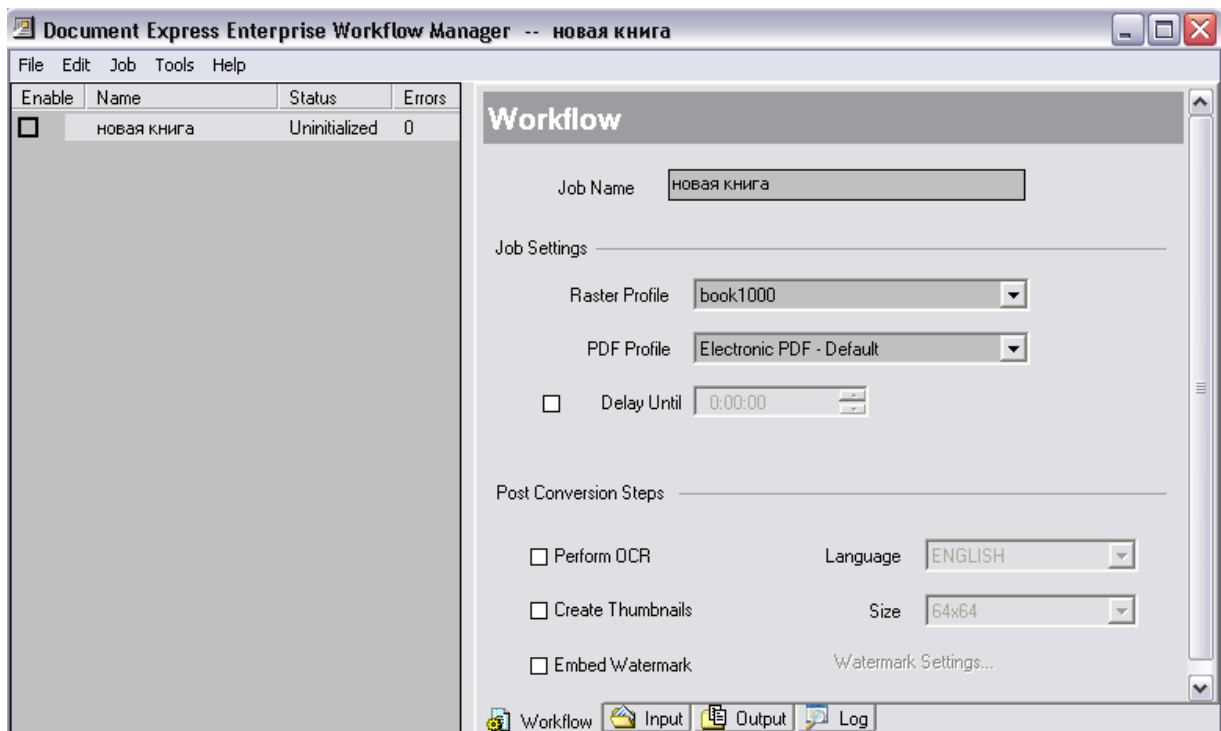
Кодировать в djvu можно двумя способами.

Первый, использовать или свободно распространяемую программу **Solo 3.1** (в этой программе используется старый алгоритм кодирования) или **Document Express Editor** версий от 4 до 6. Это делается просто, запускаем программу, загружаем первую страницу (только одну!), добавляем к первой странице остальные, но не более 500 (обычно 200÷300). Сохраняем с выбором профиля **bitonal** и с разрешением 600 dpi (для уменьшения размера выходного файла, в файле конфигурации **documenttodjvu.conf** для **Solo** или **Editor** ставим **pages-per-dict** равным не менее 100).

Второй, рекомендуемый способ, это использовать **Document Express Enterprise 5.1** (доступна облегченная версия этого пакета, объемом около 20 мегабайт). Вначале создаем профиль для кодирования (делается это не часто, можно один раз), для этого запускаем **Document Express Enterprise Configuration Manager** из этого же пакета, из списка выбираем профиль **Bitonal(600)**, нажимаем на кнопку **Advanced Settings...**, в диалоговом окне **Advanced Settings** выбираем закладку **Text** и ставим **Pages Per Dictionary** равным 1000 (конечно, это небольшой экстремизм, можно ограничиться значением 100÷200). Сохраняем этот профиль под новым именем. Увеличение размера страниц на словарь, приводит к заметному уменьшению размера файла, до 25%. Профиль **Bitonal** используется только для черно-белых сканов, если у Вас есть страницы с иллюстрациями, то в этом случае лучше использовать профиль на основу **Scaned**.



Запускаем **Document Express Enterprise Workflow Manager**, загружаем все страницы зараз, в поле **Job Name** пишем название книги, из списка **Raster Profile** выбираем, подготовленный ранее профиль, переключаемся на закладку **Output** и из списка **Separate Document(s) by** выбираем **One document only**. Ставим галочку (с самого левого края под **Enable**) и ждем конца кодирования, следим или пока эта галка исчезнет или по закладке **Log**.



5. Создание текстового слоя

После того, как все уже сделано, остановится на этом просто себя не уважать, не говоря уж об остальных.

Для добавления распознанного текстового слоя в djvu книгу потребуется две программы. Первая это **FineReader 7.0** или **8.0** версии. Вторая программа, это небольшая утилита **DjvuOCR 2.1** от болгарского камрада **Gencho**.

Загружаем все, обработанные в кромсаторе, тифы в **FineReader**, те из которых была сделана djvu книжка, и распознаем в пакетном режиме. Ура, ура, ура, появилась новая версия этой утилиты 2.2, в которой сняты эти ограничения. Теперь можно редактировать текст после распознавания в ФР, соблюдая некоторые ограничения:

- а) при редактировании сохранять некоторые символы оригинального текста (например интервалы), т.е. не переписывать большие блоки;
- б) желательно сохранять количество строк в параграфе (т.е. не стирать знаки конца строки).

Кстати, для наших целей вполне подойдет триальная версия **FineReader**, которую можно свободно скачать с официального сайта разработчика.

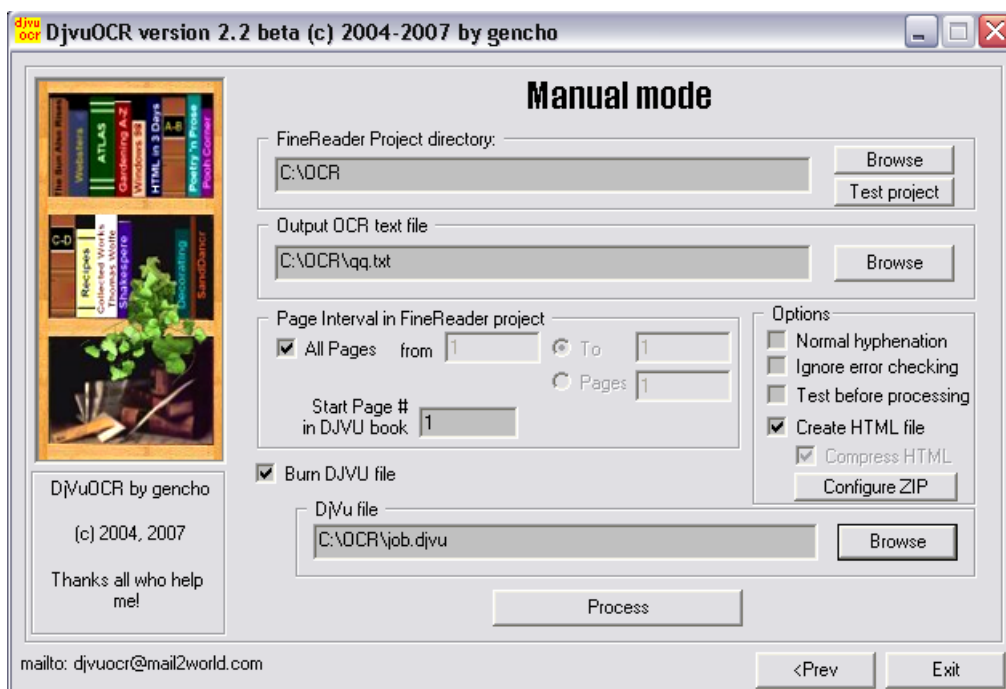
Недавно появилась программа **FineReader 9.0**, для нее разрабатывается **DjvuOCR 2.4**. Запускаем **DjvuOCR**, ждем на кнопку **Manual made OCR manager**



Далее, тоже все просто:

- **FineReader Project directory** – выбираем каталог с проектом.
- **Output OCR text file** – это любой, пустой текстовый файл, помещенный в каталог с проектом.
- Ставим галочку на **Burn DJVU file** и выбираем djvu книжку.
- Нажимаем **Process**.
- Ждем несколько минут.

И всё.



6. Добавление обложки

В добавлении обложки, если не преследовать сверхзадач, никаких особых хитростей нет. Сканируем обложку в цвете в 200 dpi, чистим ее по вкусу, слегка размываем ее и кодируем в djvu профилем Photo(300). Полученный файл добавляем в книгу, например с помощью **Document Express Editor**.

З.Ы. Поступают жалобы, что таким образом сделанная обложка имеет размер меньший, чем страницы книги, что выглядит не эстетично, поэтому делайте обложку, как Вам больше нравится, сохраняя размер в пределах разумного. Говорят, есть методики так называемого отдельного кодирования по созданию сверх компактных обложек в 600 dpi, если не лень, поищите.

7. Оглавление

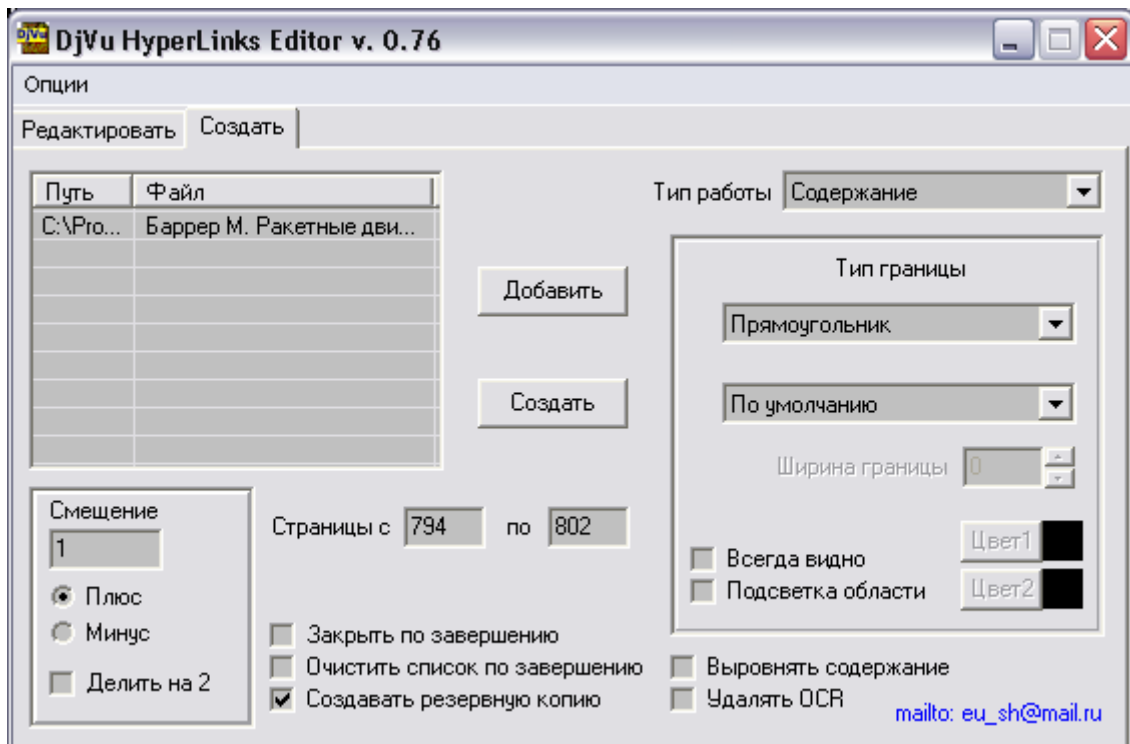
Знаете ли Вы, что в Вашу книгу можно вставить оглавление? А между прочим, благодаря уважаемому **Shea**, это поразительно просто! Для этого воспользуемся утилитой **DjVu Hyperlinks Editor**.

Добавляем книгу, указываем, на каких страницах находится оглавление (нумерация с учетом обложки), выставляем смещение 1 (для компенсации обложки) и ждем **Создать!**

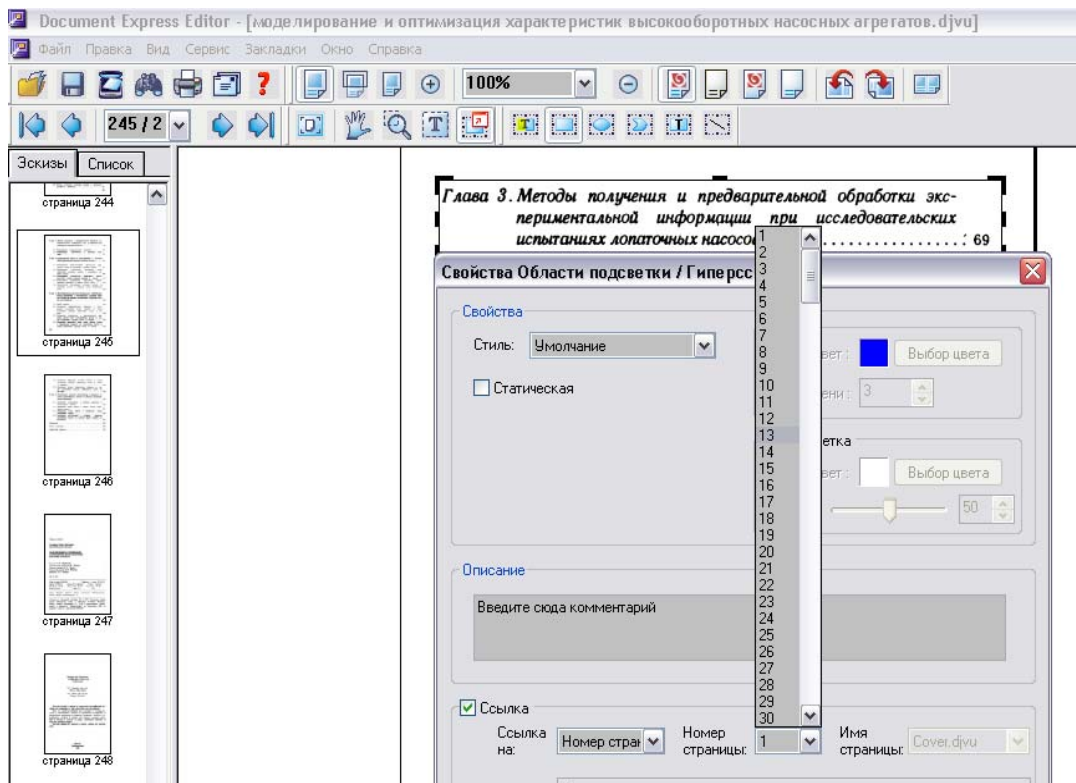
Конечно, без глюков пока не обходится, проверьте на всякий случай результат, вопиющие случаи можно и поправить в ручную (см. вставку оглавления ручным случаем чуть ниже).

Аналогично, с помощью этой программы, можно создать предметный указатель (выбираем **Тип работы**).

Подробнее, обо всех возможностях программы, можно почитать в сопроводительном файле.



В 5 и 6 версии **Document Express Editor** это же можно сделать мышкой. Жмем на кнопку – прямоугольная гиперссылка, обводим пункт меню, выскакивает окно диалога – свойство гиперссылки, в котором выбираем линк на номер страницы и затем соответственно сам этот номер. Ну и так далее, пока рука не отсохнет. Только делаем это в самый последний момент, после добавления обложки, вкладок и пр., иначе ссылки сдвинутся.



7. Используемые программы и где их взять

IrfanView	www.irfanview.com	freeware
ScanKromsator 5.9	http://www.djvu-soft.narod.ru/	freeware
Solo 3.1	http://www.djvu-soft.narod.ru/	freeware
Document Express Editor	http://www.djvu-soft.narod.ru/	?
Document Express Enterprise	http://www.djvu-soft.narod.ru/	?
ABBYY FineReader	www.abbyy.com	trial
DjvuOCR 2.4 beta	http://djvuocr.ucoz.ru/	freeware
DjVu Hyperlinks Editor	http://www.djvu-soft.narod.ru/	freeware

За <http://www.djvu-soft.narod.ru/> особое спасибо **monday2000!**

Заключение

С образцами книг, выделанных строго по этой инструкции (подчеркиваю строго!), можно ознакомиться на ... Если результат Вас удовлетворит, то может быть и сами попробуете?

Прежде чем делать книгу, проверьте, может она уже есть, посетите поисковый ресурс www.poiskknig.ru. Хотя если Вам встретится некачественный экземпляр, то никто не мешает его переделать (практически любая djvu книга, сделанная в 300 dpi ч/б и менее, может считаться браком).

Поделиться книгой, можно опубликовав ее на ...

P.S. Несколько грязных слов о ФайнРидере.

Если Вы собираетесь сделать научно-техническую книгу в формате djvu – **не надо использовать ФР для сканирования и обработки сканов!**

Примите это как данность, если Вы хотите получить качественный результат, несмотря на предлагаемые ФР удобства по типу всё-в-одном, попробуйте все же данную инструкцию.

Из основных недостатков:

- 1) использование сжатия на основе jpeg, что, как минимум, приведет к раздуванию e-книги после кодирования;
- 2) примитивно реализованный алгоритм выравнивания страницы;
- 3) если Вы сканируете в 300 dpi в градациях серого, то вся обработка будет выполнена для этого разрешения, в то время как в кромсаторе, сначала идет ресемплинг до 600.

P.S. С выходом девятки появилась возможность отключать пресловутое принудительное выравнивание страниц.

P.S. Помните, что все вышеизложенное, не есть истина в последней инстанции, просто здесь систематизирован подход, практически гарантирующий неплохой результат. Но никто не запрещает, как использовать другие методики (Фотошоп с плагинами, BookRestorer, Corel PHOTO-PAINT, RasterID..., были попытки использовать MatLab), так и экспериментировать с кромсатором (но вначале, настаиваю, сделайте несколько разворотов строго по инструкции, что бы было с чем сравнивать). Путь получения хорошего скана книги много, главное на выходе иметь 600 dpi!