# Инструкция по созданию книг в формате DjVu
# Издание третье, исправленное

DMVN Corporation

11 ноября 2006 года

## Введение

Перед тем, как сканировать какую-либо книгу, надо сначала понять, не сделано ли это до Вас кем-либо ещё. Для этого рекомендуется поискать по разным электронным библиотекам (по фамилии автора, например) это творение. Если книги и правда нигде не найдено, то можно приступать к работе.

При создании книги нужно исходить из следующих соображений. Во-первых, получаемый файл должен быть разумного размера. Так, книга объёмом 300–400 страниц не должна занимать больше 8–10 мегабайт. Во-вторых, получаемый файл DjVu-файл должен быть таким, чтобы его потом можно было напечатать, и при этом текст можно было прочесть. Хорошо, если при этом на страницах не будет содержаться ничего лишнего, то есть чёрных краёв. В этом случае при печати тонер или чернила расходуются только на текст.

Далее мы расскажем, как делать книги, удовлетворяющие этим критериям. Следует иметь в виду, что предлагаемый способ не претендует на скорость/оптимальность/универсальность/идеальное качество/что либо ещё. Это просто один из методов, дающий в целом неплохой результат.

Написанное здесь относится к книгам, которые не содержат в себе полутоновых изображений. Если Вы делаете книгу, которая их содержит, мы советуем делать так: все чёрно-белые странички сканировать по технологии, описанной здесь, а для полутоновых применять специальные ухищрения. Возможно, мы расскажем о них в следующих версиях данной инструкции.

Эта технология может показаться сложной, однако результат оправдывает себя. Чтобы не надоедало делать однотипные операции, слушайте музыку или смотрите телевизор, при этом количество механических ошибок уменьшается (человек вообще не приспособлен к выполнению однотипных операций в течение долгого времени, а это помогает ). Сейчас, когда электронные книги уже достаточно распространены, нередко приходится сталкиваться с ужасными по качеству «работами неизвестных художников», читать или печатать которые затруднительно либо бесполезно. Следует помнить, что чтение с экрана (коим приходится иногда заниматься из-за отсутствия возможности распечатать книгу) утомляет глаза, а если скан плохой, то это может приводить к быстрому утомлению глаз и т. п. Но не будем далее растекаться мыслею по древу, а приступим к работе...

## 1. Получение изображений

Существует несколько методов получения изображений. Страницы можно фотографировать или сканировать. Последний метод является наиболее качественным, поэтому мы будем описывать именно его.

Кроме того, бывает такая ситуация, что книга уже была отсканирована, но плохо. Мы поговорим о возможных способах лечения плохих изображений страниц в разделе «Исправление недостатков сканов».

Пусть пока для простоты наша книга не содержит цветных (или полутоновых) иллюстраций и цветного текста. Тогда лучше всего для сканирования использовать программу FineReader 8.0 (далее — FR для краткости).

### 1.1. Настройки сканера в FR

Некоторые производители сканеров предоставляют некоторый набор ПО для сканирования. Оно часто работает медленно и не так, как нужно. В FR имеется возможность использовать его собственный интерфейс для работы со сканером, и это обычно удобнее и быстрее. Чтобы использовать этот интерфейс, нужно в настройках сканера выбрать пункт Use FR Interface (Использовать интерфейс FR).

Важными параметрами сканирования являются *разрешение* (resolution) и *яркость* (brightness). В FR эти настройки имеются в диалоговом окне Scanner Settings (Настройки сканера). Скажем про них пару слов.

Оптическое разрешение — это то количество точек на дюйм, которое будет содержать получаемое изображение. Чем оно больше, тем больше получается файл и тем качественнее получается само изображение. Для книг приемлемым является разрешение 600 DPI (Dots Per Inch — точек на дюйм).

Что касается яркости, то очень часто хватает автоматической настройки, но иногда, если бумага слишком тёмная, или шрифт слишком жирный, её приходится регулировать вручную. Для этого в диалоге настроек яркости нужно выбрать пункт Manual (Вручную) и установить ползунок в нужное положение. Какое оно должно быть, определяется опытным путём: выберите типичную страницу, отсканируйте с разной яркостью и посмотрите, какое положение ползунка даёт наиболее качественное изображение.

При этом нужно помнить, что установленная вручную яркость может перестать быть приемлемой, если изменить разрешение.

## 1.2. Сканирование страниц

Кладите книгу на сканер как можно более ровно. Чтобы не было проблем, снимите крышку со сканера, она всё равно только мешает. Главное — прижать книгу посильнее на сгибе, чтобы страница поплотнее прилегала к стеклу (только не продавите стекло сканера!). При этом надо следить за тем, чтобы книга не смещалась (если это произошло, не поленитесь, пересканируйте страницу). Если не жалко времени — сканируйте не разворот целиком, а по одной странице, так получается качественнее. При сканировании толстых книг бывает удобно класть книгу углом, при этом вторая половина висит вертикально сбоку (вот тут держать приходится особенно крепко).

Если Вы сканируете разворот книги целиком, то надо установить ориентацию страницы в Landscape, тогда при обрезке страницы не придётся поворачивать на 90°.

После того как книга отсканирована, проверьте количество страниц! (Оценить фактическое количество страниц в книге по нумерации весьма несложно.) Если оно не совпало, то надо методом половинного деления выяснять, где произошла лажа (это позволяет за $\sim \log_2 n$ операций локализовать место).

Если с количеством всё в порядке, сохраняем *пакет* (batch). FR сохраняет все страницы в формате TIFF (чёрно-белом). При сохранении FR создает папку, названную по имени пакета, с кучей файлов (отдельных страниц), к каждому из них прилагается .frf-файл.

После этого рекомендуется скопировать пакет и работать с его копией. Дело в том, что операции FR типа обрезки страниц не являются обратимыми, поэтому, если у Вас дрогнула рука, и Вы испортили страницу, можно будет восстановить изображение из копии пакета (надо просто закрыть FR и подменить в пакете соответствующий tiff-файл).

## 1.3. Деление страниц пополам, очистка от мусора и обрезка

Обычной ситуацией является то, что каждая отсканированная страница содержит в себе две страницы книги (разворот). Чтобы с DjVu-файлом было удобно работать, надо их разделить. Для этого в FR есть специальная функция Split Image (*Разбить Изображение*). При этом FR предупреждает, что разбиение — необратимый процесс, и что оно не может быть отменено. Ставим галочку, чтобы он не показывал этого сообщения всякий раз, и после этого начинаем делить. Для ускорения процесса выбираем Add Vertical Separator(*Добавить Вертикальный Разделитель*) и потом делаем так: мышкой устанавливаем место, где следует разрубить страницу, а на клавиатуре жмём Enter всякий раз, когда разделитель установлен. При этом FR будет по умолчанию автоматически переходить к следующей странице (а нам только того и надо!).

После разделения, если мы видим, что нигде не ошиблись, можно ещё раз сохранить пакет (и удалить оригинал). Смысл в том, **чтобы перед началом очередной операции сохранять копию**, чтобы можно было в случае чего откатиться назад.

Однако при обрезке мы никак не убираем тот чёрный мусор, который остаётся на развороте. Сделать странички чистыми помогает функция Crop (*обрезка*). Она работает аналогично разделению страниц. Выделяем область, в которую влезает текст целиком и ещё немного, чтобы меленькие белые поля оставались, и жмём кнопку. При этом совершенно замечательно то, что обрезающий прямоугольник сохраняет размеры при переходе от страницы к странице, так что его остаётся только правильно отпозиционировать. Таким образом, все страницы получаются одного размера (что красиво и удобно).

В случае, если какая-то грязь попала внутрь страницы и не была убрана при обрезке, можно убрать её с помощью инструмента типа ластика. Им можно убирать прямоугольные области.

Кроме того, в FR есть специальная функция Despeckle (не знаю, как оно переведено на русский), которая убирает со страницы мелкую грязь и мусор. Чтобы применить это преобразование ко всем страницам одновременно, можно выделить их все и после этого уже выбрать этот пункт меню.

После этого сохраняем пакет ещё раз. Обычно после выполнения всех этих работ хочется передохнуть, поэтому самое время дать вслать поработать компьютеру, а самому в это время кофейку выпить. Для этого как раз подходит следующий шаг.

# 2. Распознавание

Распознавание текста (оно же Optical Character Recognition, или, для краткости, OCR) в книге позволяет потом искать в ней текст подобно тому, как это можно делать в обычных текстовых файлах и PDF-документах, что очень удобно. Это делается с помощью того же самого FR.

В отличие от предыдущих шагов, этот процесс прост. Смотрим, на каком языке написана наша книга, выбираем соответствующий язык в FR и начинаем распознавание — выбираем Read All Pages (*Распознать Все Страницы*). Кстати, для книжек по математике на русском языке всё равно имеет смысл ставить Russian/English, потому что значки и буковки тогда распознаются лучше. Распознавание — довольно долгий процесс. После его завершения опять сохраняем пакет.

Очень важен тот факт, что после распознавания ни в коем случае нельзя менять файлы изображений, потому что при изменении может измениться положение текста на странице, а это значит, что привязка координат распознанного текста будет нарушена.

Кроме того, ни в коем случае нельзя пытаться редактировать текст, полученный FR. Всё как есть, никаких опечаток не исправляем, а то будет плохо потом, при внедрении текста.

Теперь надо бы заняться кодированием изображений в DjVu. Этой нетривиальной операции и посвящён следующий шаг.

# 3. Кодирование в DjVu

## 3.1. Подготовка к кодированию

TIFF-файлы, полученные FR, имеют одну неприятную особенность — внутри них содержится preview-копия странички. Их проще всего убрать, отконвертировав все TIFF-файлы в монохромные BMP. Это можно очень легко сделать с помощью утилиты IrfanView (её можно скачать с сайта `http://www.irfanview.com`). Эта утилита умеет преобразовывать группы файлов (File|Batch Conversion/Rename...). Настройки там тривиальны, не буду их описывать, тем более, что всё сводится к указанию типа выходного файла (Bitmap) и опции Batch Conversion, а дальше надо просто выбрать список файлов для конверсии, указать путь для выходных файлов и нажать Start. После конверсии из TIFF в BMP preview-странички будут изгнаны.

## 3.2. Кодирование в DjVu

Мы будем использовать утилиту DjVu Document Express Enterprise 5.1 (не путать с Document Express Pro!). Это наиболее могущественный инструмент, поддерживающий много форматов кодирования. Для начала запускаем Configuration Manager (конфигуратор). Эта утилита позволяет выбрать и сконфигурировать *профиль кодирования* (Profile). По умолчанию уже создано несколько готовых профилей, и их по сути достаточно, но, если Вы достаточно освоились с этой программой, то можете экспериментировать с настройками и делать собственные профили.

Нас будет интересовать профиль Bitonal 600 DPI. Для простоты можно ничего не менять, а оставить все его настройки как есть, они и так достаточно хорошие. Изложим, однако, общий принцип. Разрешение должно совпадать с разрешением сканирования, а режим Bitonal довольно хорошо подходит для чисто монохромных страниц (если страница содержит полутоновые рисунки, то надо использовать Scanned или Photo, но мы такие случаи пока не рассматриваем). На самом деле возможности у этой утилиты ещё больше, просто редко когда бывает нужно разбираться со всеми режимами одновременно. Кстати, чтобы понять, на что способен каждый профиль, можно загрузить тестовую страничку и посмотреть, во что она превратится при кодировании.

После того как разбирательства с профилями завершены, закрываем конфигуратор и запускаем Workflow Manager. Эта утилита осуществляет, собственно говоря, процесс кодирования. В ней выбираем профиль (Raster Profile) (один из тех, которые были в конфигураторе), выбираем список входных файлов на вкладке Input, выбираем имя выходного файла (и то, куда его надо кидать) на вкладке Output, устанавливаем Separate Documents By в One document only, чтобы в итоге у нас получился один файл, а не много мелких (по одному на страницу).

После этого уже можно поставить галочку напротив нашего Job-а, и тогда начнётся процесс кодирования. Это может быть довольно долго и требует большой работы от процессора. При кодировании статус меняется на Running, по завершении становится Complete.

Как только кодирование завершено, и Вы убедились в том, что полученный DjVu-файл нормально открывается и всё в порядке, можно, наконец, вспомнить, что мы когда-то долго и упорно распознавали книжный текст.

### 3.3. Внедрение текста в DjVu-файл

Далее нам пригодится разработка болгарского программиста Gencho под названием DjVuOCR. Её можно скачать с сайта `http://dmvn.mexmat.net`. Эта утилита умеет преобразовывать распознанный FR-ом текст в некоторый формат, который добавляется в качестве текстового слоя в DjVu-файл. Я назвал этот процесс *внедрением*, хотя в оригинале это звучит как Burn OCR info into DjVu file. Итак, запускаем утилиту DjVuOCR (мы будем рассказывать про версию 2.1) и выбираем Manual mode OCR manager. Затем указываем путь к пакету (FR Project Directory), указываем что-нибудь типа `[book name]-ocr.txt` в строке Output OCR text file. Опцию Create HTML можно вырубить, остальное по умолчанию. Далее ставим опцию Burn DjVu file и подсовываем туда тот DjVu-файл, который мы получили в процессе кодирования. (На всякий случай рекомендуем перед этим процессом делать резервную копию самого DjVu-файла, чтобы его случайно не запороть. Глюков за этой программой пока не замечено, но предосторожность не повредит.)

После этого давим магическую кнопку Process, и через минуту DjVu-файл готов — в него внедрён текст. После этого рекомендуется открыть файл какой-нибудь программой, умеющей читать DjVu, и проверить, что текст таки ищется. Вот и всё! Поздравляем Вас с созданием очередной электронной книги!

В этой утилите был обнаружен один неприятный дефект (был обнаружен благодаря Светлане Балагезян, которая попыталась делать книги по предыдущей инструкции). Именно, если пакет назван именем с русскими символами или в имени содержатся пробелы, то программа может глючить и выдавать всякие непонятные сообщения об ошибках типа FR project does not exist. Чтобы этого не происходило, кладите пакет в корень диска и называйте его попроще (используйте только английские буквы и не более 8 символов).

## 4. Исправление недостатков сканов

А этот раздел мы напишем чуть позже... когда у нас опять появится время.

## Послесловие

Если у Вас после прочтения этой инструкции возникли вопросы, или есть желание что-либо добавить или посоветовать, милости просим. Пишите нам в почту `dmvn@mccme.ru` (или ICQ 244633817), обновления данной инструкции — на сайте Научной Библиотеки МехМата МГУ (`http://lib.mexmat.ru`).

Если Вы испытываете трудности с добыванием того или иного программного обеспечения, обращайтесь в почту `dmvn@mccme.ru`, не стесняйтесь, поможем чем сможем. Часть программных продуктов действительно очень сложно найти (речь идёт не о демо-версиях, а о полных версиях с ключами).

А вообще DMVN в Сети живёт вот тут: `http://dmvn.mexmat.net`.