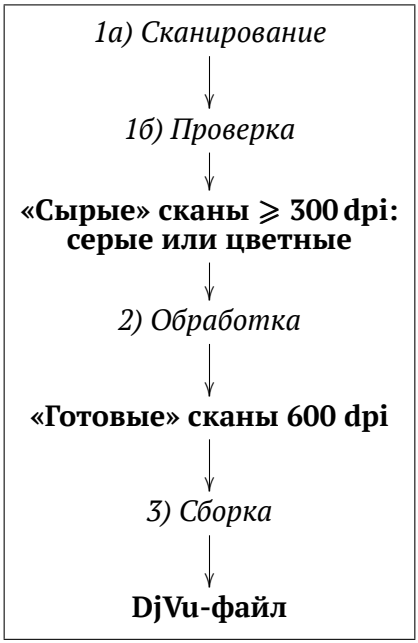
Алексей Шипунов

КАК СКАНИРОВАТЬ

Очень короткое руководство

Версия 16 марта 2021 г.



А. Шипунов. Как сканировать. Очень короткое руководство. Версия 16 марта 2021 г.

Рекомендации, основанные на опыте создания онлайн-библиотеки «Флора и фауна».

Оглавление

Предисловие

Я отсканировал, обработал и собрал в электронную книгу приблизительно 20% библиотеки «Флора и фауна». Это больше двух тысяч книг. Еще примерно 30% библиотеки—это книги, в которых я так или иначе принял участие. Например, обработал чьи-то сканы, и/или собрал электронную книгу для выкладки на сайт. Я думаю, это достаточно большой опыт, чтобы им можно было поделиться. Вот почему я написал этот текст.

Важный момент: цель «Флоры и фауны» — это предоставить книги *для работы*. Цель «сделать идеально» или «сделать так, чтобы можно было забыть о бумажном варианте» не ставилась. Есть сообщества сканировщиков, которые эти цели преследуют, и возможно, читателю нужно обратиться туда и не читать дальнейшее.

С другой стороны, научными электронными книгами, как мне кажется, должно быть удобно пользоваться, то есть нужно, чтобы их можно было разыскать в сети, быстро скачать, найти нужное, скопировать информацию (в том числе иллюстрации) для дальнейшего использования, а если требуется (скажем, для полевой работы) — то и распечатать с максимальной скоростью. Эти принципы были основными при организации библиотеки.

Самый быстрый способ

Я решил сделать эту маленькую главу для тех, кто не хочет читать дальше. Вот как сделать электронную книгу всего за четыре ступени:

1. Отсканировать книгу на сканере с разрешением 300 точек на дюйм, и при этом поставить вывод в тонах серого или цветной (но не черно-белый!)
2. Проверить сканы, чтобы были на месте все страницы и чтобы все части текста были резкими и без теней.
3. Сохранить файлы в формат JPEG, или при помощи любой программы-конвертора (скажем, IrfanView) перевести их в этот формат.
4. Запустить утилиту j2p.exe (Windows)[1] или img2pdf (другие операционные системы). Эти утилиты делают PDF без каких-либо изменений в исходных сканах. Да, и **ни в коем случае** не пытаться уменьшить размер получающегося файла!

Дальше можно этот файл хранить, читать с экрана, выкладывать, делиться, или посылать (тому, кто прочитал эту книжку дальше[2]) для дальнейшей обработки.

[1] Доступна здесь: https://sourceforge.net/projects/jpgtopdf/.

[2] Для тех, кому прислали этот файл: вынуть обратно из такого PDF сканы лучше всего при помощи утилиты mutool — см. дальше по тексту.

Глава 1

Как сканировать

1.1. Три главных принципа

1. Сканировать надо с разрешением не меньше **300** dpi.
2. Сканы должны быть «серыми» или цветными.
3. Для того, чтобы сделать электронную книгу, сканы надо обязательно **обработать** специальными программами (например, Scan Tailor).

1.2. … и один полезный совет

1. Необработанные файлы сканов нужно хранить — чем дольше, тем лучше.

1.3. Сканер

1. Годится любой, но есть сканеры (типа Canon LiDE), которые очень чувствительны к расстоянию между бумагой и стеклом.
2. Надо очень внимательно следить за неровностями бумаги, особенно изгибом у корешка. Потом никакие программы Вам не помогут, если на скане вместо текста — размытые пятна.

3. Сохранять изображения надо в TIFF 300 dpi со сжатием LZW (каждое слово здесь значимо).[1]

4. Черно-белые и серые страницы сканировать в тонах серого (grayscale / greyscale), а обложки, цветные вклейки и т.п. сканировать в цвете.

5. Карты и крупные вклейки сканировать по кускам, потом придется их соединять в специализированном панорамном софте или просто в GIMP/фотошопе, лучше всего — перед обработкой.[2]

1.4. Как не пропускать страницы

Это большая беда, потому что пропущенная страница обязательно кому-то потребуется, а книга, как правило, уже далеко. Поэтому:

1. Проверять надо обязательно «не отходя от кассы», сразу после сканирования.

2. Считать все страницы долго и нудно, поэтому я проверяю «девятками»:

   а) загружаю все отсканированное в быстро работающую программу-просмотрщик графики (скажем, IrfanView);

   б) быстро листаю, говоря вслух или про себя «один, три, пять, семь, девять» и одновременно смотрю на то место, где находится номер *нечетной* страницы (скажем, нижний правый угол);

   в) если вдруг есть несовпадение, возвращаюсь назад и листаю медленно, разбираясь, что к чему.

[1] В принципе, можно сохранять и в других форматах, например, в JPEG от 80% качества, но необходимо увериться, что изображение при сжатии не страдает.

[2] Или не соединять. Я в последнее время перестал это делать, но, что интересно, пользоваться несклеенными картами вполне удобно (таблицами — менее удобно, но терпимо).

3. Частая ошибка — пропуск первых страниц, причем методом счета она отлавливается с трудом. Поэтому надо обязательно еще раз проверить, все ли начальные страницы на месте.
4. Да, самое главное: никогда не сканируйте книгу частями! Пропущенное Вами, скорее всего, нужно кому-нибудь другому.

1.5. Фотоаппарат вместо сканера

1. Фотографировать только по одной странице (не разворотами).[3]
2. Всегда использовать максимальный возможный размер кадра.
3. Нужны минимум два источника света (скажем, лампа и окно). Хотя фотографировать просто на подоконнике — тоже неплохо, а еще лучше получается на улице в пасмурную погоду.
4. Держать фотоаппарат как можно параллельнее плоскости книги.
5. Обязательно проверить резкость всех кадров и *сразу же* заменить негодные.
6. Фотографировать книги в твердой обложке гораздо удобнее.

Учтите, что лист A4 для 300 dpi требует 18 мегапикселей (к счастью, бóльшая часть книг имеет страницы меньшего размера, скажем, для книги малого формата может быть достаточно и 6 мегапикселей).

Главный враг фотографирования книг — неровное освещение, а уже потом темнота и геометрические искажения. Нужно обеспечить *как можно более ровный свет*. Еще раз: свет — самое главное!

[3] Я часто нарушаю это правило для экономии времени. Однако надо убедиться, что разрешение получившихся кадров не меньше 300 dpi.

1.6. Еще о фотографировании книг

Фотографировать книги трудно физически: надо как-то держать и книгу, и камеру. Что я делал:

1. Держал книгу одной рукой и фотографировал другой. Работает с маленькой камерой (скажем, Canon S100), но в большинстве случаев не очень удобно. Можно дополнительно предварительно разглаживать страницы. Минусов немало, в частности, надо следить, чтобы пальцы не въезжали в область текста (пусть даже пустую на данной странице).

2. Клал толстую книгу (книги) на края разворота либо на верх его. Минус — каждый раз надо приподнимать.

3. Просил кого-то держать развороты. Очень эффективно и ускоряет процесс в разы, минус очевиден: нужен второй человек. И за пальцами надо следить.

4. Подкладывал книгу или чехол камеры под ту часть разворота, где меньше страниц. Не решает всех проблем, но помогает. Многие старые книги (пока еще типографское качество было высокое, где-то до середины 1960-х) и не требуют больше никаких дополнительных усилий.

5. Тонкие, тесно сшитые книги переворачивал в последней трети процесса на 180°. Так было гораздо удобнее их прижимать. Ну а потом, перед обработкой, эти кадры программно переворачивал.

1.7. Нумерация сканов

Иногда это очень важно, потому что утилиты, собирающие из сканов DjVu или PDF, могут делать это неправильно, и тогда нарушается порядок страниц.

1. Не нужно использовать в названиях файлов русские буквы, буквы в верхнем регистре и пробелы. Только цифры, латин-

ские буквы в нижнем регистре, подчеркивание, ну и еще дефис.

2. Файлы надо нумеровать с лидирующими нулями. То есть первый файл скана может называться `sc-001.jpg`, второй файл — `sc-002.jpg`, а последний, скажем, `sc-239.jpg`. Все обложки, форзацы, вклейки тоже должны быть в этом ряду. Тогда проблем с неправильной сборкой не будет.

3. Сделать такое можно, например, с помощью Total Commander. А под Linux у меня есть такой скрипт:

```
index=1
prefix=$1
for file in `ls -v *$2`; do
    ext="${file##*.}"
    counter=`printf %03d $index`
    newname=$prefix-$counter.$ext
    echo "Renaming $file to $newname"
    mv $file $newname
    index=$(($index+1))
done
```

Глава 2

Как обрабатывать

2.1. Обработка

1. Собственно говоря, есть только одна вменяемая программа обработки: **Scan Tailor**[1].
2. Перед обработкой нужно обязательно узнать разрешение исходных сканов (как — см. ниже).
3. После выделения «полезных зон» проверить все страницы на предмет отрезанных кусков текста и лишнего белого места.
4. На этапе полей уменьшить поля по умолчанию, и исключить страницы с очень большой шириной или высотой из выравнивания (нужно воспользоваться **сортировкой по размеру**[2]). Если этого не сделать, поля будут слишком велики.
5. На этапе вывода обязательно выводить в 600 dpi.

2.2. Как узнать разрешение

1. Очень рекомендую поставить быстрый просмотрщик изображений, например, IrfanView.

[1]... и множество менее вменяемых, например, ScanKromsator, unpaper, gscan2pdf и др.

[2]Переключатель находится в правом нижнем углу экрана.

2. Надо открыть файл и посмотреть, какое разрешение записано в файле (форматы TIFF и JPEG хранят информацию о разрешении), и если что-то не так — проверить при помощи правила 6,5 строчек.

3. **Правило 6,5 строчек**: высота 6–7 строчек средней книги (в пикселах) *примерно* равна разрешению. Вот как это делается в IrfanView (Рис. 1). У скана разрешение 600 dpi, потому что высота 6,5 строчек *примерно* равна 600 пикселам.

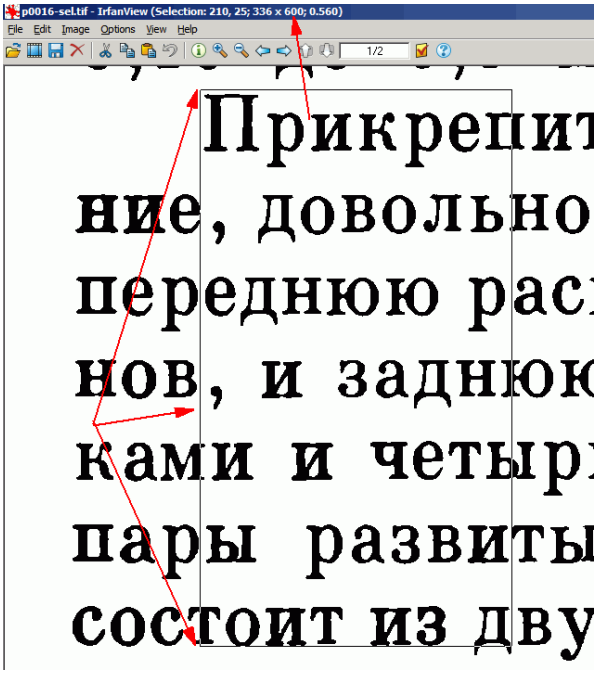


Рис. 1. Правило 6,5 строчек. Разрешение здесь 600 dpi, потому что высота 6,5 строчек *примерно* равна 600 пикселам (размер в пикселах см. в заголовке окна программы).

2.3. Как начать работать со Scan Tailor

Надо скачать и установить программу Scan Tailor Plus или любую другую разновидность Scan Tailor, умеющую работать с *прямоугольными зонами картинок* (например, Scan Tailor Featured, Scan Tailor Universal).

Вот короткая инструкция по пользованию: http://djvu-soft.narod.ru/st_quick.htm, ниже см. измененный мной вариант.

Вот инструкция подлиннее: http://djvu-soft.narod.ru/st_ref.htm

2.3.1. Инструкция по работе со Scan Tailor

1. Создать проект, импортировать туда все нужные файлы.
2. Если надо, то вручную поправить ориентацию страниц на стадии «Исправление ориентации».
3. Перейти к разрезке страниц, начиная с самой первой. Через несколько страниц станет ясно, какой тип самый частый. Его надо установить его для всех последующих страниц. Затем перейти в конец и проверить, правильно ли разрезались страницы там.
4. **Перейти** на стадию «**Полезная область**», проверить в ручном режиме первые несколько страниц, а потом запустить пакетную обработку и ждать, пока она завершится.
5. После этого *просмотреть каждую страницу* в нормальном размере (а не только в ленте справа), отлавливая случаи, когда какие-то части текста (чаще всего номера страниц) *отрезаны*. Если такое случилось, надо поправить полезную область или перейти на одну из предыдущих стадий и поправить там. Хорошо бы поправить и случаи, когда в полезную область попало лишнее.

Это самая долгая, но очень важная стадия. Иногда, если сканы очень грязные, можно начать со стадии сортировки (см. ниже), но затем нужно опять вернуться сюда.

6. В процессе просмотра можно увеличить полезную область в тех местах, где страницы начинаются *не с самого верха* (в качестве подсказки можно использовать видимый иногда на сканах оборот страницы). Это поможет на стадии выравнивания («Макет страницы»).

7. Потом два раза (по ширине и по высоте) отсортировать по размеру полезной обрасти, перейти вверх и вниз ленты и каждый раз проверить самые большие и самые маленькие (особенно самые большие). Здесь я обычно убираю также типографские пометки для сшивки тетрадок.

8. **Перейти** на стадию «**Макет страницы**», и сразу уменьшить поля до 5 мм со всех сторон и применить их ко всем страницам. Потом отсортировать по ширине и, начиная с самой широкой, убрать из выравнивания те страницы, которые значительно отличаются от других (например, обложка, некоторые таблицы). У обложки лучше убрать вообще все поля.

9. Потом переходить вверх по списку страниц, обращая внимание на поля. Если какие-то страницы по-прежнему сильно шире остальных и вызывают увеличение полей во всей книге, убрать их из выравнивания. То же самое проделать со списком страниц, выравненным по высоте.

10. Потом перейти в начало списка страниц, выравненных по высоте и *центрировать* некоторые очень короткие страницы (скажем, титульные). Потом отсортировать по порядку и посмотреть, нет ли еще страниц, которые лучше выровнять по центру (чаще всего они находятся ближе к началу или концу книги).

11. **Перейти** на стадию «**Вывод**» и вывести первые несколько страниц, тем самым подобрать подходящие параметры вывода, черноты и очистки, которые затем применить ко всем остальным страницам.

12. Потом заняться исключениями: страницами с отточиями, обложкой, цветными страницами, страницами с фотографиями. В каждом случае нужно, как правило, применить свои параметры:

    а) у страниц с отточиями поменять параметры очистки мусора, иначе точки могут потеряться;

    б) для прямоугольных фотографий надо выбрать смешанный вывод, а там выбрать прямоугольные зоны картинок;

    в) обложку и вставки, почти целиком заполненные полутоновыми или цветными картинками, сделать цветными (поля у обложки не нужны, а в остальных случаях надо решать на месте).

13. Теперь опять запустить пакетную обработку. Файлы будут записаны в директорию, указанную при создании проекта (по умолчанию в поддиректорию `out`).

14. После этого (а можно и до начала вывода) нужно почистить страницы от мусора. Для этого есть вкладка *зон заливки*, где надо выделять и заливать мусор белым.

    Редкий мусор (скажем, печать на 17-ой странице) можно удалять сразу на этапе просмотра полезной области, переходя на вывод. Не страшно, если потом изменятся поля, потому что программа сохраняет абсолютные координаты области заливки.

15. В конце работы надо обязательно **просмотреть** результат, по крайней мере в ленте, и поправить ошибки. Особенно часто возникают ошибки с определением страниц с полутоновыми рисунками (иногда трудно на взгляд отличить штриховые или точковые от полутоновых), и еще с выравниванием. Первые обычно видны как «кляксы».

2.4. Обработка перефотографированных книг

Самое серьезное отличие (хороших) фотографий от сканов в том, что они все немного разного размера. Отсюда особенности обработки.

1. Все страницы с черно-белыми (полутоновыми, серыми) фото переконвертировать в grayscale / greyscale.
2. Запустить Scan Tailor таким же образом, как для обычных сканов, но сразу же вручную выставить разрешение (см. выше про правило 6,5 строчек). Разрешение можно выставить и заранее, в файлах фотографий.
3. На этапе полезной области постараться вручную подравнять по остальным короткие (последние в главах, например) и узкие (некоторые рисунки и таблицы) страницы.
4. На этапе макета страницы присвоить всем одинаковые поля и *снять выравнивание*.
5. На этом же этапе выровнять друг по другу 2–3 самые короткие и 2–3 самые узкие страницы. Можно и больше, чем 2–3.
6. На этапе вывода попробовать задействовать автоматическое выправление искривлений (ориентировочно с первой страницы, заполненной текстом и до оглавления).
7. Выправление искривлений очень полезно именно для перефотографированных страниц, но оно плохо работает на страницах с узкими колонками текста, коротких страницах и страницах, начинающихся с картинок (и наоборот, хорошо работает, если на страницах есть верхняя горизонтальная линия).

   По опыту самое простое — сделать временный DjVu и «пройтись» по нему, отлавливая плохо обработанные страницы. Обращайте особенное внимание на низы страниц, потому что алгоритм выправления иногда теряет куски текста.

   После этого надо запустить вывод еще раз и сделать окончательный DjVu.

Альтернативный метод — сделать два вывода: с выправлением искривлений и без, потом просмотрщиком изображений удалить плохо выправленное и копированием поверх добавить невыправленное.

Как видно, возни больше, чем со сканами, а результат, увы, всегда хуже. Более того, на правильной передаче *цвета* в большинстве случаев можно поставить крест. Но если ничего другого, кроме как сфотографировать книгу на подоконнике не остается, то надо учесть все, что написано выше.

2.5. Если книга цветная

1. Обычно (в научной литературе) цветная — только обложка. Заднюю обложку я по большей части игнорирую, а переднюю вывожу из Scan Tailor безо всякой дополнительной обработки и полей, только с повышенным обычным образом (600 dpi) разрешением. Так же можно поступать с цветными и серыми вклейками, где нет или мало текста.

2. Если есть и рисунки (цветные или серые), и текст, то их обработку надо разделить. Это называется *сегментация*. Тут спасают прямоугольные области изображения на этапе вывода. Надо выбрать смешанный режим, и там — форму области. Прямоугольные области доступны в нескольких разновидностях программы Scan Tailor.[3] После окончания вывода надо еще раз пройтись по всем таким страницам, проверить и подправить (если надо) границы зон.

3. Если есть цветной фон или цветной текст, то это караул. Есть отдельные достижения, но в целом никакая программа обработки не умеет их как следует обрабатывать. Проще всего вообще не разбивать такую страницу на текст и рисунки — выводить как обложку. Либо, на свой страх и риск, конвертировать в черно-белые.[4]

[3] Например, Scan Tailor Plus, Scan Tailor Featured, Scan Tailor Universal.

[4] Впрочем, можно попробовать перевести такие страницы в *малоцветные* (*постеризовать*) и потом закодировать специальным DjVu-методом (`cpaldjvu`). Это

4. При конвертировании получившегося в DjVu есть три варианта:

   а) довериться программе, пусть сама решит как ужимать (*автоматическое сегментирование*);

   б) указать программе, где рисунки, а где текст (*вставка картинок*);

   в) заставить программу считать все страницы картинками (то есть обойтись без сегментации).

   Более подробно об этих вариантах написано в последней главе.

2.6. Поля, штампы и последние страницы

1. Мне представляется, что исторически большие поля служили для того, чтобы:

   а) было можно обрезать книжный блок несколько раз при последующем переплете (многие книги выпускались непереплетенными и даже неразрезанными);

   б) не пачкать пальцы и бумагу, перелистывая страницы.

   Обе причины ничего общего с электронными книгами не имеют. Поэтому в Scan Tailor я всегда ставлю маленькие поля (5×5 мм), чтобы было удобнее читать с экрана. А чтобы избежать увеличения полей за счет выравнивания, выключаю из выравнивания самые длинные и самые широкие страницы (Scan Tailor позволяет сортировать по этим признакам на этапе выравнивания).

2. Библиотечные штампы, посвящения и прочие «послепечатные» дополнения я тоже удаляю. В Scan Tailor на этапе вывода есть вкладка «Зоны заливки» («Fill zones»), которая работа-

умеет, например, моя программа `img2djvu`. С ее помощью я иногда делаю малоцветные обложки.

ет как ластик, позволяя залить любым цветом произвольную область страницы.

3. Последние страницы с рекламой других книг, страницами «для заметок» и задней обложкой я чаще всего не включаю в книгу.

2.7. Поворачивать ли страницы

1. Это в основном карты и таблицы, иногда рисунки или фотографии.
2. Scan Tailor не позволяет повернуть страницы на 90° перед выводом, а если повернуть вначале, то возникают проблемы с выравниванием или повернуть вначале вообще нельзя (потому что чаще всего книга отсканирована разворотами, а нужная страница — только часть разворота).
3. С другой стороны, большинство просмотрщиков DjVu/PDF позволяют на ходу повернуть любую страницу. А если просматривать по две страницы сразу, то заранее повернутые страницы создают проблемы.
4. Но с еще одной стороны, разпознавалки текста лучше работают с заранее правильно повернутыми таблицами, да и кодировщик DjVu делает файлы меньших размеров, если страницы правильно ориентированы.
5. Наконец, с недавних пор утилита djvused из набора DjVu Libre позволяет просто поставить индикатор поворота прямо в DjVu-файл при помощи команды `set-rotation`. Такой способ вообще не требует никаких серьезных усилий.

В общем, решайте сами.

2.8. Если сканер может только черно-белое

Это бывает у старых ксероксов-сканеров-принтеров. Они могут очень быстро работать (тысячи страниц в час), но не умеют сканировать цветное или в тонах серого.

1. Тогда обработка сканов при помощи увеличения с бинаризацией практически бессмысленна. Нужно просто запустить пакетный конвертор изображений (тот же IrfanView).
2. В IrfanView надо выбрать скан из середины книги, и примерно прикинуть при помощи выделения, как обрезать края. В заголовке окна будет виден размер обрезки и координаты левой верхней точки выделения.
3. На том же изображении перейти в пакетный режим, выбрать все сканы одной книги.
4. Установить параметры обрезки (значения взять из заголовка окна, где все еще есть выделение) и запустить пакетную обрезку.
5. После обрезки пролистать сканы и проверить, не обрезалось ли чего лишнего. Если да, то удалить неправильно обрезанное, изменить границы и повторить процесс (с книгами в мягкой обложке приходится иногда до пяти раз повторять).
6. Несколько первых страниц, страницу с печатью (обычно 17-ую) и задние страницы загрузить в GIMP/Photoshop и дополнительно почистить.

Глава 3

Как собрать электронную книгу

3.1. DjVu или PDF

1. Из DjVu очень легко сделать PDF, наоборот — гораздо труднее.
2. Сделать DjVu гораздо труднее, чем сделать PDF, причем некачественный и не поддающийся исправлению DjVu сделать легче качественного.
3. DjVu всегда будет в 2–5 раз компактнее PDF (если PDF собран из сканов), а при одинаковом размере — будет в несколько раз быстрее листаться и занимать меньше оперативной памяти.
4. DjVu не «портит изображения», если кто-то так думает, то он просто не умеет его готовить ©

3.2. Чем делать DjVu

1. DjVu Small.
2. DjVu Toy (программа еще кучу всего умеет делать).
3. Поскольку обе вышеуказанные программы — это обертки над консольными утилитами, можно напрямую запускать их ядро — программу `documenttodjvu.exe` (этот файл можно вытащить из DjVu Small).

4. Поскольку лицензионность вышеописанных программ сомнительна, можно воспользоваться свободной библиотекой DjVu Libre (http://djvu.sourceforge.net/).

   Под Linux я сделал работающий «комбайн», который многое умеет: https://github.com/ashipunov/img2djvu. Чтобы в нем сделать из папки вывода Scan Tailor (она называется `out`) DjVu с «вклеенными» иллюстрациями, можно набрать:

   ```
   $ img2djvu -d 600 -a 2 -l 2 out
   ```

5. **Не нужно делать** факсимильные электронные книги[1] при помощи FineReader любых версий, а также программ, которые ставят штампы на страницы.

3.3. Как конвертировать в DjVu сканы с картинками

Этот способ рассчитан на вывод программы Scan Tailor и разрешение 600 dpi.

1. Скачать DjVu Small http://djvu-soft.narod.ru/scan/djvu_small.htm и внимательно прочитать инструкцию.

2. Найти там в папке программы файл `documenttodjvu.conf`, и модифицировать его, добавив в конец следующие строчки:

   ```
   #@displayName:user B/W (600 dpi)
   my_600: very-aggressive300
   dpi=-600
   pages-per-dict=40

   #@displayName:600 with images very aggressive
   BW600_with_images_very_aggressive: my_600
   threshold-level=100
   ```

[1] То есть книги, состоящие из сканированных страниц, а не из распознанного и заново сверстанного текста.

```
shape-filter-level=100
bg-subsample=1
fg-subsample=2
quality=75
pages-per-dict=20
```

3. После этого становятся доступны два новых профиля. Если не черно-белая только обложка, то лучше всего использовать «user B/W (600 dpi)», а если много фото, то «600 with images very aggressive».

   Этот второй профиль был опробован на множестве книг с иллюстрациями. Он, в сравнении с остальными вариантами, делает страницы весьма высокого качества.

Глава 4

Маленькие хитрости

Здесь я решил собрать всякие дополнительные советы, относящиеся к книгосканированию. Я не хочу загромождать текст ссылками, просто погуглите названия, этого должно быть достаточно.

Большинство приведенных ниже программ рассчитаны на работу в консоли (терминале, командной строке). Настоятельно рекомендую этому научиться! Такое умение упрощает и ускоряет работу во множество раз.

4.1. Как все-таки сделать DjVu из PDF

Допустим, что у вас есть файл `my.pdf`, где страницы совершенно не обработаны, то есть изображения со сканера просто подшиты в файл. Такой файл обычно очень велик по размеру, с трудом листается, и плохо поддается распознаванию. Чтобы сделать из него DjVu, надо сначала «вынуть» все страницы, превратив их в изображения:

1. Если картинки со сканера вставлены целиком, то можно попробовать запустить в консоли утилиту pdfimages из набора xPDF Tools:

   ```
   pdfimages my.pdf 1
   ```

   Получаются PPM-файлы, надо обязательно определить их разрешение и потом конвертировать в TIFF или JPEG. Если серые картинки были отсканированы как цветные, то очень полез-

но перевести их обратно в тона серого, чтобы исчез желтоватый или синеватый оттенок.

2. Если PDF сделан просто, то очень быстро можно вынуть картинки при помощи MuPDF:

   ```
   mutool extract my.pdf
   ```

3. Формат PDF — это монстр. Ваш файл может оказаться нестандартным и поэтому не поддающимся pdfimages и mutool. Например, вместо картинок будут выведены их маленькие фрагменты. Тогда нужно выяснить разрешение (правило 6,5 строчек) и запустить утилиту pdftoppm (из тех же xPDF Tools), задав ей найденное разрешение. Кроме того, pdftoppm по умолчанию создает PPM-файлы, они очень велики, поэтому можно задать ей опцию сохранять все в TIFF или JPEG:

   ```
   pdftoppm -r 300 -jpeg my.pdf 1
   ```

4. Вместо pdftoppm можно попробовать Ghostscript. Команда для конвертации файла такая:

   ```
   gs -q -dBATCH -dNOPAUSE -sDEVICE=jpeg -r300 \
    -dTextAlphaBits=4 -dGraphicsAlphaBits=4 \
    -sOutputFile=%04d.jpg my.pdf
   ```

   Можно еще попробовать «mutool draw», но она работает сильно медленнее и не умеет выводить картинки в TIFF или JPEG.

Ну а после того как страницы-картинки получены, можно перейти ко второй главе и заняться сканообработкой.

Есть и альтернативный вариант, но он предназначен не для «PDF со сканера», а для очень больших «PDF из типографии», где просто хочется уменьшить размер файла и ускорить с ним работу. Для этого нужно поставить утилиту pdf2djvu:

```
pdf2djvu -d 600 -j 4 my.pdf -o my.djvu
```

4.2. Как распознать текст

Тут речь пойдет не о том, как распознать текст вообще, а о том, как сделать DjVu файл с текстовым слоем. Такой слой очень полезен, например, для поиска. А можно скопировать текст и вставить его в Google Translate для перевода. Можно, наконец, вынуть весь текст и превратить DjVu в EPUB, FB2 или другой формат текстовой электронной книги (правда, это требует большой работы, которая почти не поддается автоматизации).

Минус такого подхода в том, что распознанный текст трудно (хотя и возможно) редактировать. Мне это не кажется большой проблемой. Через несколько лет технологии распознавания, скорее всего, улучшатся, и проще будет еще раз распознать текст, вместо того чтобы тратить время на его тщательную вычитку.

Самое простое (для тех, конечно, кто умеет работать с консолью) — это запустить утилиту ocrodjvu:

```
ocrodjvu --in-place -l rus+lat my.djvu
```

Чтобы распознавание прошло нормально, должна быть установлена какая-нибудь OCR-программа, по умолчанию подразумевается, что это tesseract. Качество распознавания у него очень высокое, единственная проблема — он работает медленно. Скорость можно увеличить, запустив процесс в несколько потоков при помощи опции «-j».

4.3. Полезные утилиты

Здесь я хочу описать несколько полезных свободных утилит, которые могут помочь на разных этапах рабочего процесса.

ImageMagick Фактически, это такой «фотошоп в командной строке», причем в отличие от оригинального фотошопа, очень легко автоматизирует любые однородные операции. Например:

```
mogrify -auto-level -selective-blur 0x1+20% \
```

```
-gaussian-blur 0x.5 -sigmoidal-contrast 5 \
-monitor -path new_directory -format jpg *.jpg
```

Длинная команда, зато она автоматически ко всем JPEG файлам из текущей директории применит авто-уровень, избирательное размытие, Гауссово размытие, S-контрастирование, запишет результат в формате JPEG в «new_directory» и покажет все операции на терминале. Кстати, я написал такую команду, чтобы обрабатывать цветные файлы, которые приходили с большого сканера Konica Minolta.

ImageMagick умеет конвертировать изображения в PDF, например

```
convert *.jpg -density 300 my.pdf
```

сделает PDF из всех JPEG файлов текущей папки (только нужно обязательно правильно определить разрешение).

GraphicsMagick Аналог ImageMagick, но работает гораздо быстрее (хотя имеет меньше возможностей).

pdftk Очень удобен для того, чтобы быстро добавить, удалить или переставить страницы, или слить PDF-файлы. Например,

```
pdftk my.pdf cat 2-end output my2.pdf
```

удалит первую страницу из PDF.

Чтобы слить два PDF, нужна такая команда:

```
pdftk 1.pdf 2.pdf cat output 12.pdf
```

С другими операциями над PDF (скажем, обрезкой страниц, вставкой текста или перемещением объектов по странице) дело гораздо сложнее. Есть много коммерческих утилит, но свободных практически не существует, причем в конце 2019 года их стало еще меньше, чем несколько лет назад. Фактически, остался лишь pdftk.

djvupages Это тоже консольная утилита, которую я написал как аналог описанных выше PDF-утилит, но для DjVu файлов. Например,

```
djvupages --delete -f 1 -l 1 my.djvu
```

удалит первую страницу из DjVu-файла. Нужно только сначала установить утилиты DjVuLibre.

Чтобы превратить весь DjVu в изображения, нужно запустить команду

```
djvupages --images my.djvu
```

djvm Чтобы слить два DjVu, лучше воспользоваться утилитой djvm из DjVuLibre напрямую:

```
djvm -c 12.djvu 1.djvu 2.djvu
```

Чтобы вставить одну DjVu-страницу внутрь другого DjVu (например, как вторую страницу), опять же лучше использовать djvm:

```
djvm -i 1.djvu 2.djvu 2
```