



На этой странице приведена самая необходимая информация о работе с электронной литературой. Более подробную информацию по всем без исключения разделам, а также смежным вопросам, можно найти в Интернете

[На главную страницу JURASSIC.RU](#)

## [Загрузить бесплатные программы для просмотра файлов \\*.PDF и \\*.DJVU:](#)

Для файлов \*.PDF: [Adobe Reader 7.0.5](#) [русская версия; 21 Мб], [Adobe Reader 8.1.1](#) [22 Мб] (*Более старые версии Adobe Reader можно найти на сайте разработчика [www.adobe.com](http://www.adobe.com)*) или [Foxit Reader 2.2](#) [для Windows; 2.1 Мб]

Для файлов \*.DJVU: [WinDJView 0.5](#) [0.5 Мб] (<http://windjview.sourceforge.net/>) или [DJVU Browser Plug-in](#) [6.6Мб] ([www.lizardtech.com](http://www.lizardtech.com)).  
Версии этих программ под Mac OS, Linux, UNIX ищите на сайтах разработчиков.

## [Основное](#)

[Почему электронный формат ?](#)

[Основные форматы электронных книг – PDF и DJVU](#)

[DJVU или PDF?](#)

[Распознавание электронных книг \(OCR\)](#)

[Как читать электронные книги на незнакомых языках](#)

[Поиск отдельных слов в тексте](#)

[Конвертация между форматами PDF и DJVU](#)

[Оцифровка бумажных книг](#)

[Какой сканер использовать?](#)

[Где достать необходимые программы?](#)

[Ничего не получается, есть важные дополнения или остались вопросы?](#)

## [Почему электронный формат ?](#)

Хранение книг и статей в электронном виде не требует большого количества свободного пространства в квартире/лаборатории, не мешает окружающим и позволяет с удобством работать одновременно с большим количеством источников, а кроме того:

- 1) позволяет с легкостью получать неограниченное число копий
- 2) дает возможность размещения в Интернете – универсальной, доступной любому человеку творческой среде
- 3) расширяет и упрощает обмен литературой
- 4) дает возможность поиска фрагментов текста по ключевым словам (если текст распознан) или по закладкам (если таковые сделаны)
- 5) дает возможность читать и понимать книги на незнакомых языках
- 6) позволяет взять с собой в командировку полноценную библиотеку

подробнее:

**М.А. Рогов, А.П. Ипполитов, М.В. Полякова.** Электронные библиотеки в Интернете и их роль для палеонтологии и стратиграфии: текущее состояние дел и перспективы дальнейшего развития. Доклад на совещании "Палеострат-2009", г. Москва, Палеонтологический институт им. А.А. Борисяка РАН, 26.01.2009 г.



[Загрузить презентацию в формате \\*.PPT с речевым сопровождением](#) [17 Mb]



[Загрузить обычную презентацию](#) [6.5 Mb]

Размещение аудиопрезентаций - дело нетипичное, и мы будем вам очень признательны, если вы оставите свое мнение / комментарии / пожелания по этому вопросу. Нужны ли подобные аудиопрезентации и стоит ли практиковать регулярное размещение такого рода материалов, разрабатывать эту идею? Можно принять участие в [опросе на Форуме](#) или высказаться в [Гостевой книге](#).

## [Основные форматы электронных книг – PDF и DJVU](#)

Все публикации, представленные на сайте, оцифрованы участниками проекта и многочисленными друзьями сайта (всем им большое спасибо!) и представлены в виде файлов с расширениями .PDF и DJVU.

### [PDF –](#)

файловый формат, первоначально созданный для обеспечения совместимости при передаче электронных документов между разными операционными системами. Наиболее распространенный формат для хранения любых электронных документов как с векторной (в том числе текстом), так и растровой графикой.

Основной инструмент для работы с PDF - программа *Adobe Acrobat* модификаций *Standard* или *Professional*, для просмотра достаточно модификации *Reader*

Замеченные недостатки:

- (-) многостраничные растровые документы (в первую очередь объемные книги), полученные путем сканирования, и **содержащие распознанный (см. ниже) векторный текстовый слой**, могут листаться с заметной задержкой (это зависит только от того КАК этот файл был распознан)
- (-) как и большинство других программных продуктов компании Adobe, *Acrobat* (самая распространенная программа для работы с PDF-файлами) не отличается компактными размерами и излишне скромными требованиями к мощности компьютера. Как следствие, на многих компьютерах работает медленно, на некоторых - очень медленно.

Достоинства:

- (+) все остальное

### [DJVU –](#)

это растровый формат, изначально предназначенный для хранения отсканированных книг и их сильного сжатия за счет разложения изображения на "текст" и "фон", которые сохраняются с различными параметрами.

Наиболее удобный инструмент для редактирования DJVU-файлов - программный пакет *Lizardtech Document Express Enterprise Edition*, а для просмотра - свободно распространяемая программа [WinDJView](#).

Достоинства:

- (+) небольшой размер файлов. Файл в формате DJVU по размеру в 1,5-10 раз меньше файла с аналогичными характеристиками растрового изображения в формате PDF (в зависимости от способа кодировки PDF и характеристик изображения)

Явные недостатки:

- (-) для кодирования в DJVU категорически не подходят цветные изображения с большим количеством важных деталей (в т.ч. карты) и рисунки (в т.ч. фототаблицы) в оттенках серого. В принципе, их можно сохранять как «*Photo*» - фон с высоким разрешением, без автоматического разложения на "фон" и "текст", но размер получаемого файла сводит на нет основное преимущество DJVU – компактность, а кроме того:
- (-) почти все программы для просмотра файлов DJVU, кроме *WinDJView*, заметно «притормаживают» на страницах, которые закодированы как фон с высоким разрешением (с опцией «*Photo*»), без разложения на "текст" и "фон".
- (-) отсутствие единой удобной программы для создания, редактирования, распознавания и просмотра файлов DJVU, из-за чего приходится устанавливать набор самостоятельных программ. При этом их суммарные возможности настроек печати, редактирования, комментирования текста, сохранения, конвертации, безопасности и пр. в них очень сильно ограничены по сравнению с *Adobe Acrobat* - основным инструментом для работы с PDF.

Кроме того, уже сегодня существуют алгоритмы, кодирующие файлы PDF способом, аналогичным для DJVU-кодировки – текст и фон отдельно с различными параметрами. Такую систему использует, например, программа, поставляемая вместе со сканерами Canon – *CanonScan Toolbox*. Преобразование таких PDF-файлов в DJVU сокращает размер максимум в 1,5-2 раза, что является нецелесообразным. Кроме того, в последние версии *Acrobat Professional* встроены инструменты, способные существенно уменьшить размер PDF-файла (уменьшение путём установки совместимости с поздними версиями программы, оптимизация и др.). Кроме того, быстрый рост мощностей компьютеров и скоростей доступа в Интернет постепенно сводит на нет «преимущество малого размера» формата DJVU.

## [DJVU или PDF ?](#)

Из сказанного выше следует, что DJVU имеет смысл использовать ТОЛЬКО для толстых книг и только в следующих случаях:

- книга не содержит большого количества цветных иллюстраций или фототаблиц в оттенках серого. Кодировка этих элементов в DJVU по алгоритму разложения на "текст" и "фон" очень резко снижает качество изображений, а сохранение в фотокачестве не дает никакого выигрыша в размерах.
- текст напечатан на какой-либо цветном или сером полутоновом фоне, детали которого выполняют декоративную функцию и неважны для читающего.
- текст в книге набран цветным шрифтом, при этом цветных иллюстраций и фототаблиц нет или их исчезающе мало.

В 90 % остальных случаев предпочтительнее оцифровка в PDF, при необходимости с дальнейшей оптимизацией полученного файла.

## [Распознавание электронных книг \(OCR\)](#)

Оптическое распознавание текста (OCR – Optical Character Recognition) - это процесс, заключающийся в преобразовании графической информации, например, полученной со сканера, в текстовый вид, пригодный для обработки текстовым редактором. Распознавание электронных книг производится при помощи специальных программ.

Поскольку точность в этом нелегком деле никогда не бывает стопроцентной, в распознанном тексте всегда присутствуют неправильно опознанные знаки, а шрифты компьютера могут сильно отличаться от шрифтов типографии - книга будет выглядеть не так же, как на бумаге ! Поэтому наиболее удобным способом хранения распознанного текста в отсканированных электронных книгах является на



сегодняшний день **создание невидимого текстового слоя** под основным растровым изображением. В этом случае пользователь на экране видит отсканированное изображение, однако к нему привязан текст, который можно **выделять и копировать** в текстовые редакторы, при необходимости исправляя вручную ошибки OCR-программы. Оба формата - и PDF, и DJVU - поддерживают такую возможность.

**Какие программы использовать для распознавания? В конце данного раздела вы можете найти результаты проведенного нами сравнительного теста для разных OCR-программ. Перечислим некоторые из них:**

#### Для формата PDF:

1) **ABBYY Fine Reader** – самая популярная OCR-программа в России.

Плюсы *Fine Reader*:

(+) очень высокая точность распознавания;

(+) высокая скорость распознавания;

(+) можно выбрать несколько языков, которые будут распознаваться. Это наиболее удобно в случае, если присутствуют «неправильные» буквы вроде немецких с умляутами, разнообразных чешских-польских и т.д. Если правильно выставить языки, шансы корректного распознавания заметно возрастают. При распознавании палеонтологических русскоязычных работ удобнее всего ставить один из «стандартных» вариантов языков распознавания файнридера – Russian-English.

Ещё один момент: иногда *Acrobat* категорически отказывается распознавать крупные работы на русском языке, «вылетая» в процессе работы. *FineReader* более устойчив: пусть неспешно, но дело он своё сделает. Заметим, что в *FineReader* можно загружать как отдельные страницы (уже на выходе сохраняя их в виде pdf-файла), так и нераспознанные pdf-ы (в последнем случае при большом объёме файла только загрузка без распознавания может занять немало времени).

Минусы:

(-) программа часто дает сбой при обработке больших книг (более 400 стр.). Вероятно, это зависит лишь от мощности компьютера

(-) после распознавания и сохранения в pdf падает качество исходной картинке: буквы в строчках слегка "приплясывают", а строчки местами "плывут" за счет появления горизонтальных и вертикальных сдвигов изображения на 1-2 пиксела.

(-) Так как программа для загрузке PDF-файлов переводит страницы в растровую форму, определяя разрешение автоматически, то общий вид книги "на выходе" получается хуже, чем "на входе". Все книги, посканированные с разрешением 400 dpi, программа растеризует как 300 dpi - с неизбежной потерей качества. Повлиять на это нельзя.

Два последних недостатка являются ключевыми при работе с PDF-файлами - качество картинке после распознавания не должно ухудшаться. **Эти минусы полностью устранены в последней версии программы, ABBYY FineReader 9.0.** Более низкие версии подходят для простого извлечения текста для разового перевода, а для создания постоянного невидимого текстового слоя – категорически не рекомендуются. Однако при отсутствии 8-й версии *Acrobat Professional*, необходимости работы с русскоязычными pdf-файлами, или необходимости распознавания книги в оттенках серого - вполне сгодится. В настройках сохранения для формата pdf должно быть обязательно выставлено, что текст сохраняется под изображением страницы. Здесь есть еще одна хитрость. После сохранения PDF из *Fine Reader 7* во всех случаях полученный файл необходимо оптимизировать программой *Adobe Acrobat* (меню *Advanced* - опция *PDF Optimizer*), немного утерев в качестве - иначе страницы будут листаться с заметной задержкой!

2) **ReadIris** - еще одна мощная программа, менее известная у нас в стране, чем *FineReader*. Хотя настройки этой программы и количество поддерживаемых языков слабо отличаются от *FineReader*, распознавание этой программой вдвое медленнее, а качество распознавания - намного хуже. Плюсом можно считать то, что программа не портит исходную картинку.

3) **Adobe Acrobat** модификации *Standard* или *Professional*. Эта программа имеет самостоятельный OCR-модуль, судя по особенностям распознавания, основанный на движке компании IRIS (см. п. 2), но с меньшим количеством языков, невозможностью выбора нескольких языков. Русский понимает только начиная с 8 версии. Минусом можно считать то, что в разных версиях эта распознавалка «прячется» в разных частях меню (это, впрочем, относится и к функциям уменьшения размера файлов). В 8й версии это *Document – OCR Text Recognition - Recognize using OCR*. Здесь нужно выставить страницы (одну или весь документ) и язык (если поставить в качестве основного языка русский, слова, написанные латиницей – те же названия окаменелостей, например - прекрасно распознаются)

Опция *"PDF output style"* определяет, что вы хотите получить на выходе. Варианты такие. *"Searchable image"* - невидимый текстовый слой под изображением (учтите, что само изображение будет новым, полученным при загрузке страницы! Ввиду этого к использованию НЕ рекомендуется - в полученном файле страницы будут листаться с заметной задержкой). *"Searchable image (Exact)"* - то же самое, но с сохранением всех первоначальных характеристик изображения (РЕКОМЕНДУЕТСЯ для создания невидимого текстового слоя). *"Formatted Text & Graphics"* - полная замена растрового изображения распознанным векторным, при этом резко сокращается размер, а все слова, в которых программа "сомневается", можно оставить в нетронутым растровом виде.

(-) мощный минус OCR-модуля программы *Adobe Acrobat* - низкое качество распознавания страниц, отсканированных в цвете или оттенках серого: между буквами в словах появляется огромное количество лишних пробелов, как следствие - текст с таких страниц невозможно перевести на другой язык, предварительно не убрав лишние пробелы вручную, кроме того, совершенно невозможно автоматически найти ключевые слова.

(-) программа работает очень медленно и отнимает большую часть ресурсов компьютера, поэтому для распознавания крупных книг лучше запускать ее тогда, когда вам компьютер не нужен - например, ночью.

(-) по невыясненным причинам при распознавании **отдельных страниц** русскоязычных книг иногда возникают ошибки. Достаточно редко - что-то около 1 "сбойной" страницы на 200-500 нормальных. Как бороться - пока неясно (даже перегонка из Акробата в bmp/tiff и обратно не помогает). Что делать: при возникновении ошибки запомнить номер "сбойной" страницы, и еще раз распознать остальные страницы в книге, кроме "сбойной", выставив нужные диапазоны страниц в диалоговом меню опции *"Recognize text using OCR"*.

**Резюме.** Для страниц, отсканированных в цвете или оттенках серого, предпочтительной программой для распознавания на сегодня является *Fine Reader*, а для черно-белых - *FineReader версии не ниже 9.0*, или, в крайнем случае, *Adobe Acrobat*. Будем надеяться, что в новых версиях *Acrobat* недостатки, связанные с распознаванием страниц в цвете и оттенках серого будут устранены.

#### Для формата DJVU:

4) *Document Express Editor* (коммерческие версии) или пакет *Document Express Enterprise Edition* целиком - он включает в себя. Качество - так себе. В программе *Document Express Editor* OCR запускается опцией *"OCR"* из меню *"Tools"*, а вот настройки распознавания (язык) прячутся в другом меню - *Edit - Preferences* - закладка *OCR*. "Взломанные" версии не понимают буквы "г" и, вероятно, "ц", безжалостно вырезают из текста знаки тире, по-видимому, поголовно принимая их за перенос. Точность распознавания средняя.

5) Если вы сканируете по типу 1 страница = 1 файл (см. ниже раздел «**Сканирование**»), полученные файлы можно распознать программой *ABBYY FineReader*, после чего сами файлы можно объединить в DJVU (как - см. ниже раздел «**Сканирование**»), а результат распознавания (из файлов в формате \*.FRF) внедрить в DJVU-файл программой *DjvuOCR* (официальный сайт программы <http://djvuocr.ucoz.ru/>). С помощью этой же программы можно легко раскодировать уже готовый DJVU в набор картинок, пригодных для распознавания ФайнРидером.

Кроме того, с помощью программ *DjvuOCR* и *Fine Reader* на раз-два можно распознать уже готовый djvu-файл. Все детали можно вычитать из помощи к программе *DjvuOCR*.

Первый способ - проще и быстрее, но второй предпочтительнее, т.к. намного точнее, хотя и несколько более трудоемкий.

6) вроде бы, открывать DJVU-файлы для распознавания умеет описанная выше *ReadIris*, но у нас эта прога стабильно дает сбой при попытке открыть DJVU. Впрочем, вряд ли стоит ожидать, что характеристики распознавания будут лучше, чем для формата PDF, а это означает, что **наиболее приемлемым способом для обработки DJVU является описанный в п. 5.**

\*\*\*\*\*

В заключение - результаты сравнительного тестирования нескольких OCR-программ на отдельно взятом компьютере. Выводы налицо.

*Исходный файл – PDF, 10 стр., ч/б, 15×24см, 300 dpi, текст на русском с включениями латыни (видовые названия фауны), 4 ч/б картинки с крапом, размер файла 787 Кб.*

*Аналогичный опыт со страницами в оттенках серого мы даже не проводили - во всех программах, кроме FineReader, функционирует движок IRIS, о качестве распознавании которого "серых" страниц применительно к Adobe Acrobat было сказано выше, и качество результата FineReader окажется намного лучше, чем у конкурентов. В пятой колонке, в отличие от первых четырех, исходным и конечным файлом является DJVU, а не PDF, и его сравнение по многим параметрам представляется некорректным, но мы его приводим для общего сравнения.*

		Adobe Acrobat 8	ABBYY Fine Reader 9	ABBYY Fine Reader 7	ReadIris 11	DJVU Document Express Editor
Затраченное время, секунды	время загрузки PDF	-	55	20	12	-
	время распознавания	269	113	117	239	21
	время сохранения PDF	1	16	2	10	-
	<b>Итого</b>	<b>270</b>	<b>184</b>	<b>139</b>	<b>261</b>	<b>21</b>
Итоговый PDF файл; распознанный текст в виде невидимого слоя	Размер конечного файла	880 Кб (+12% исх.)	885 Кб (+12% исх.)	843 Кб (+7% исх.)	817 Кб (+4% исх.)	-
	Качество изображения (при параметрах сохранения, соответствующих параметрам исходного файла)	<b>исходное</b>	<b>= исходное</b>	<b>1) искажения со сдвигами в 1-2 пиксел по горизонтали и вертикали</b> <b>2) изменение размеров исходных страниц (приращение полей), причем для разных страниц – разное, без видимой закономерности</b>	<b>= исходное</b>	<b>= исходное</b>
Качество автоматич. распознавания сложных картинок	Картинки узнаны?	да	Да	Да	да	да
	Текст внутри а картинок распознан?	<b>частично</b>	<b>Частично</b>	<b>почти нет</b>	<b>Частично</b>	<b>Частично</b>
	Распознаны ли крапы как текст?	<b>местами</b>	<b>Почти нет</b>	<b>Местами</b>	<b>Местами</b>	<b>Местами</b>
Фрагмент 1 – обчный русскоязычный текст хорошего качества (544 знака с пробелами)	Попытки определить возраст морсковской толщи сводятся к привязке ее по острокам и спорово-пыльцевым комплексам к разрезам восточных районов платформы и Урала, датированным по гониатитам, кораллам и брахиоподам. В зависимости от интерпретации возраста предполагаемых аналогов морсковской толщи определяется и возраст последней. Единства в решении этого вопроса пока нет. Так, например, по А. А. Рождественской, комплексы остроков из морсковских и мосоловских слов Центральных районов аналогичны комплексам верхнебийских и афонинских слов восточных районов Русской платформы, содержащих элементы нижеживетской фауны.	Попытки определить возраст морсковской толщи сводятся к привязке ее по острокам и спорово-пыльцевым комплексам к разрезам восточных районов платформы и Урала, датированным по гониатитам, кораллам и брахиоподам. В зависимости от интерпретации возраста предполагаемых аналогов морсковской толщи определяется и возраст последней. Единства в решении этого вопроса пока нет. Так, например, по А. А. Рождественской, комплексы остроков из морсковских и мосоловских слов Центральных районов аналогичны комплексам верхнебийских и афонинских слов восточных районов Русской платформы, содержащих элементы нижеживетской фауны.	Попытки определить возраст морсковской толщи сводятся к привязке ее по острокам и спорово-пыльцевым комплексам к разрезам восточных районов платформы и Урала, датированным по гониатитам, кораллам и брахиоподам. В зависимости от интерпретации возраста предполагаемых аналогов морсковской толщи определяется и возраст последней. Единства в решении этого вопроса пока нет. Так, например, по А. А. Рождественской, комплексы остроков из морсковских и мосоловских слов Центральных районов аналогичны комплексам верхнебийских и афонинских слов восточных районов	Попытки определить возраст морсковской толщи сводятся к привязке ее по острокам и спорово-пыльцевым комплексам к разрезам восточных районов платформы и Урала, датированным по гониатитам, кораллам и брахиоподам. В зависимости от интерпретации возраста предполагаемых аналогов морсковской толщи определяется и возраст последней. Единства в решении этого вопроса пока нет. Так, например, по А. А. Рождественской, комплексы остроков из морсковских и мосоловских слов Центральных районов аналогичны комплексам верхнебийских и афонинских слов восточных районов Русской	Попытки определить возраст морсковской толщи сводятся к привязке ее по острокам и спорово-пыльцевым комплексам к разрезам восточных районов платформы и Урала, датированным по гониатитам, кораллам и брахиоподам. В зависимости от интерпретации возраста предполагаемых аналогов морсковской толщи определяется и возраст последней. Единства в решении этого вопроса пока нет. Так, например, по А. А. Рождественской, комплексы остроков из морсковских и мосоловских слов Центральных районов аналогичны комплексам верхнебийских и афонинских слов восточных районов Русской	Попытки определить возраст морсковской толщи сводятся к привязке ее по острокам и спорово-пыльцевым комплексам к разрезам восточных районов платформы и Урала, датированным по гониатитам, кораллам и брахиоподам. В зависимости от интерпретации возраста предполагаемых аналогов морсковской толщи определяется и возраст последней. Единства в решении этого вопроса пока нет. Так, например, по А. А. Рождественской, комплексы остроков из морсковских и мосоловских слов Центральных районов аналогичны комплексам верхнебийских и афонинских слов восточных районов Русской







2) [www.translate.ru](http://www.translate.ru) – онлайн-новая версия программы *Promt*. Для использования требуется подключение к Интернету. Кроме перечисленных в п. 1. направлений перевода, умеет переводить с португальского на английский

3) [www.translate.google.ru](http://www.translate.google.ru) – еще один онлайн-переводчик, качество похуже, чем у *Промта*, но направлений перевода больше. Кроме перевода текстов в нём имеется также система перевода целых веб-сайтов - надо только вбить в нужное окно адрес интересующего сайта.

Для языков, с которыми не работают Гугл и Промт, в Интернете можно найти другие онлайн-переводчики.

Конечно, вряд ли вы получите удовольствие от чтения романа, переведенного *Промтом* или *Гуглом*, но вот со смыслом небольших статей или абзацев - разберетесь. Кроме того, весьма полезным в таком деле является и использование словарей, как установленных на компьютер (типа *ABBY Lingvo*), так и онлайн-овых (таких, напр, как <http://multitrans.ru/>). В отличие от программ-переводчиков в словаре можно узнать разные значения слов (переводчики выбирают обычно самый распространенный – но зачастую далеко не самый правильный перевод слова)

## Поиск отдельных слов в тексте

Еще один неповторимое достоинство электронных книг. Допустим, вам нужно найти в книге в несколько сотен страниц упоминание какого-либо вида ископаемых.... В акробате функция поиска похожа на ту, что встроена в *MS Word*: нажимаем *Edit – Find* (или просто *Ctrl+F*) – и щелкаем от одного до другого места появления нужного слова. Есть ещё один вариант – *Edit-Search*. В таком случае все упоминания вылезают в виде ссылок на страницы и можно сразу же из большого числа выбрать нужное.... все! ДЖВЮшные программы работают аналогичным образом, в них тоже есть команда *Find*, которая прячется в меню *Navigate, Edit* или *View*. Для поиска текст, правда, должен быть **РАСПОЗНАН**. И ещё. Поскольку корректность распознавания текста зависит от различных параметров (как он был пропечатан, с каким качеством отсканирован, какой язык выбран для распознавания...), часто лучше искать не по целым словам, а по их частям. Пусть найдётся что-то ненужное – но при этом можно быть более-менее уверенным, что всё, что нужно – тоже нашлось

## Конвертация между форматами PDF и DJVU (На случай, если вы - ортодоксальный приверженец одного из них)

### DJVU → PDF:

Конечный файл PDF обычно получается в 4-6 раз больше исходного, поэтому людей, пользующихся DJVU, вопрос о конвертации в PDF волнует мало. Поскольку программы для работы с DJVU не умеют сохранять файлы ни в каких форматах, кроме DJVU ;), основная возможность здесь - распечатка документа DJVU на виртуальном принтере PDF. Для этого, открыв любой ДеЖаВЮшной программой файл, отправьте его на печать, и в диалоговом окне меню *Print* выберите не тот принтер, который Вы обычно используете для печати на бумаге, а тот, который называется "*Adobe PDF*" (скорее всего, он уже был установлен ранее вместе с программой *Adobe Acrobat*). Потребуется ввести название конечного файла и выбрать папку для его сохранения. Все!

**Учтите, что для сохранения качества изображений основные опции печати на виртуальном принтере (разрешение печати - dpi и цветовой режим - Bitmap/Grayscale/Color) - должны совпадать с характеристиками отсканированного изображения (представление о разрешении можно получить, например, щелкнув по странице правой кнопкой мыши, далее "Page information"). Если разные страницы посканированы с разными опциями - печатать их придется отдельно.**

### PDF → DJVU:

Наиболее удобными являются следующие способы:

- 1) автоматическая конвертация с помощью программы *pdf2djvu*, входящей в состав пакета *Document Express Enterprise Edition*, или ее производных. Достоинства – возможность обработать сразу много файлов (причем можно сразу по завершении конвертации распознать файлы), составить повременной план на все операции (используя *DJVU Workflow Manager*) и запустить программу, например, для работы ночью, оставив компьютер включенным. Недостатки – по неясным причинам программа в процессе работы часто отказывается конвертировать отдельные файлы, реже - отдельные страницы. Часто подвешивает компьютер, быстро и с успехом заполняя оперативную память компьютера (даже при установке самого низкого приоритета для процесса!). Работает медленно – за исключением PDFов с черно-белыми изображениями, обработка вручную в разы, если не десятки раз, быстрее. Кроме того, не предусмотрена такая чрезвычайно необходимая опция как «ПАУЗА»... Если книга содержит отдельные страницы в цвете или оттенках серого, их все равно придется конвертировать вручную (см. п. 3) - иначе либо безвозвратно утеряется качество изображений, либо не уменьшится (а даже прирастет) размер.
- 2) автоматическая конвертация с помощью виртуального принтера *LizardTech Virtual Printer* (входит в состав пакета *Document Express Enterprise*). Работает аналогично принтеру *Adobe PDF* (см. выше, в начале раздела). Недостатки те же, что и при использовании программы *pdf2djvu* (п. 1), только сбоев еще больше, а явных достоинств не замечено. Грустно наблюдать, как принтер пытается конвертировать PDF с большим количеством страниц (больше 20)
- 3) Конвертация вручную при помощи программы *Document Express Editor*. Этот способ является предпочтительным, и, как ни парадоксально, он намного быстрее автоматической конвертации. Файл PDF открывается Акробатом, далее «Сохранить как», в опциях выберите «TIFF», сохраните – получится набор отдельных страниц в формате TIFF, которые затем объединяются в DJVU либо автоматически с помощью программы *DJVU Workflow Manager* (входит в состав пакета *Document Express Enterprise*), либо с использованием опции «Insert pages» в программе *Document Express Editor* и последующим сохранением файла. Второе удобнее, т.к. позволяет сохранить разные страницы с разными параметрами, что удобно, когда есть страницы с рисунками в цвете или оттенках серого. За один раз можно добавлять от 1 до 60 страниц, выделив необходимое количество.

## Оцифровка бумажных книг

### Необходимое оборудование и программное обеспечение:

1) сканер

2) программы для редактирования PDF (*Adobe Acrobat*, желательна модификация *Standard* или *Professional*, а не *Reader*) или DJVU (*Document Express Editor* или хотя бы *DJVU Solo 3.1*) - в зависимости от желаемого формата

В любой формат текст сканировать следует с разрешением не менее 300 dpi (оптимум - 400, иногда 600 dpi) ! Для картинок в цвете или оттенках серого достаточно 200 dpi, но лучше использовать 300-400 dpi - потерянное качество никогда не вернется! Конечно, при плохом качестве печати с низким разрешением сканирование изображения с высоким разрешением ничего не даст.

### Сканирование в PDF :

**Оптимальные опции для сканирования разных типов страниц: 1) Черно-белый текст - 400 dpi; Bitmap (он же BW - Black and White) (в Grayscale - только в крайнем случае - слишком сильно возрастает размер файлов!); 2) Ч/б текст + фотографии в оттенках серого или цветные - 400 dpi; Grayscale / Color, соответственно; 3) только фотографии в оттенках серого или цветные + , возможно, текст, не имеющий большого смыслового значения (надписи типа "Таблица 10", служебные пометки) - 200-300 dpi, Grayscale / Color, соответственно. Разные типы страниц лучше всего отсканировать отдельно, а затем объединить их и расставить в правильном порядке с помощью мышки и программы Adobe Acrobat (вкладка Pages)**

Существует два альтернативных алгоритма оцифровки книг в данный формат - **А)** очень простой и **Б)** не очень простой.

Первый (**А**) используется людьми, для которых содержание книги намного важнее ее внешнего облика и которые не готовы тратить очень много лишнего времени на процесс оцифровки. Тут есть два пути, зависящих от типа используемого Вами сканера.

1) Некоторые модели (*Canon Lide*, последние *Epson* и, вероятно, некоторые другие) позволяют напрямую сохранять отсканированные файлы в многостраничные pdf'ы, причем сразу же можно задать их **распознавание**. Плюс очевиден – серьезная экономия времени. Минус – если работа на русском, её придётся распознавать заново (встроенные в сканеры программы автоматически распознают западноевропейские языки, а некоторые новые модели (такие как *Canon Lide 600F*) - и русский, но при этом качество распознавания оставляет желать лучшего и такие работы есть смысл распознавать заново).

2) Для сканеров, сохраняющих отсканированное постранично, хорошим подспорьем является использование программ для работы с изображениями вроде *IrfanView* или *XnView*. Рассмотрим это на примере *XnView* (дистрибутив можно загрузить [здесь](#) [9,47 Мб]): *Файл – Сканировать в –* выбираем папку, задаём название файлов, первоначальный номер, формат (предпочтительно – .TIFF) – и нажимаем «Сканировать». Дальше остаётся переворачивать страницы и нажимать мышкой (или клавишей Enter) на соответствующую кнопку в программе сканера (не забудьте для начала сделать предварительный просмотр страниц (*Preview*), чтобы не отсканировать лишнее, выставить тип сканирования (ч/б, в оттенках серого или цветное) а по возможности – сразу задать степень яркости/контрастности (большинство сканеров это позволяют сделать). А если вы собираетесь этим путем сканировать много, то, наверное, стоит потратить некоторое время на освоение программы [Scan Kromsator](#).

Далее итоговый файл PDF собирается "одним щелчком мыши" с помощью опции *File - Combine Files* программы *Adobe Acrobat*. Есть еще более простой путь. После установки Акробата обычно можно выделять файлы прямо при просмотре папок проводником Windows, и, кликнув правой кнопкой мыши, использовать опцию "*Combine supported files in Acrobat*".

В полученном файле с помощью *Adobe Acrobat* можно переставлять страницы в нужном порядке, поворачивать их на 90/180 градусов, если они были посканированы не как надо, обрезать (точнее, скрыть) белые поля страницы по краям.

### Сканирование в формат DJVU

**Оптимальные опции для сканирования и сохранения разных типов страниц. 1) черно-белый текст - сканирование 400 dpi, Bitmap, опции сохранения 400 dpi, Normal/Bitonal, 2) ч/б текст + фотографии в оттенках серого, детали которых не имеют большого значения (обнажения, панорамы и т.п. - не палеонтологические таблицы!): сканирование 400 dpi, Grayscale, сохранение 400 dpi, Bitonal (не Normal!). 3) ч/б текст + фотографии в оттенках серого с важными деталями: сканирование 400 dpi, Grayscale, сохранение - 400 dpi, Photo. 4) только изображения в оттенках серого: сканирование 300 dpi, Grayscale, сохранение - 300 dpi, Photo. 5) цветной текст, схемы без полутоновых переходов - сканирование 400 dpi, Color, сохранение 400 dpi, Normal/Drawing. 6) детальные цветные фото - сканирование 300 dpi, Color, сохранение 300 dpi, Photo. Опция сохранения "Text quality" во всех без исключения случаях - quasilossless/lossless.**

Сканирование в DJVU сходно со сканированием в PDF. Оно также может осуществляться двумя способами.

1) использование опции *File – Scan Pages* в программе *Document Express Editor* или *File – Acquire* в *DJVU Solo 3.1*. Программа автоматически запустит привычный вам драйвер сканера, после сканирования страницы снова используете ту же опцию и в появившемся окошке выбираете «*Scan to current document*». Операцию повторяете нужное число раз. В зависимости от мощности компьютера нужно сохранять файл через каждые 10-100 страниц, а также каждый раз, когда вы хотите изменить опции сохранения изображения. Посканировали сотню страниц текста – сохраняем с опциями "*Bitmap*" или "*Normal*", 400 dpi, затем посканировали страницу с фотографией – опять сохраняемся, но теперь с опциями "*Photo*" и разрешением 200-300 dpi, затем можно вновь сканировать текст.

2) сканировать страницы в растровый формат (например, \*.TIFF), а затем объединить их с помощью *WorkFlow Manager* пакета *Document Express Enterprise* или в программе *Document Express Editor* открыть первый файл (*File-Open*; только в опции «Тип файлов» укажите «All supported image files»), затем добавить к нему другие страницы с помощью опции *Edit-Insert page after*, затем сохранить документ. С помощью опции *Insert page after* за один раз можно добавлять не более 60 файлов.

Б) Люди, для которых облик книги значит не меньше, чем ее содержание, готовы ради этого тратить много времени на изготовление цифровых копий, и используют **алгоритм Б**. Нет предела совершенству! Подробное описание основных операций можно найти здесь: <http://chemister.da.ru/Other/Text/convert.htm> (только не обращайтесь внимания на фразу «сканирование книги является противозаконным действием» - к научной литературе, сканируемой для некоммерческого использования, это не относится;) или здесь: [http://djvu-soft.narod.ru/b\\_cr.htm](http://djvu-soft.narod.ru/b_cr.htm)

Дополнительную информацию о работе с djvu см. здесь: [http://djvu-soft.narod.ru/scan/scan\\_and\\_share\\_1\\_07.htm](http://djvu-soft.narod.ru/scan/scan_and_share_1_07.htm)

Несколько слов можно добавить относительно обработки страниц с помощью программы [Adobe Photoshop](#) (если Вы желаете получить на выходе качественный файл). Первоначальное сканирование лучше всего выполнять в оттенках серого даже для текста и ч/б схем. В дальнейшем открываем нужные файлы, что нужно – поворачиваем, что нужно – стираем (всякие черные полосы по краям страниц и им подобные ненужности), а потом дружно делаем страницу пригодной для перевода в *bitmap* (ч/б формат) – и переводим. К счастью, в *Photoshop* есть замечательный инструмент, позволяющий один раз задать нужную программу действий, а потом повторять её нажатием одной единственной кнопки. Называется это чудо «*Действия*» или «*Операции*» (в зависимости от перевода) или «*Actions*» (найти его можно во вкладке «*Окна*» (*Windows*)). Нажимаем на «*New action*» (видел при этом меню, очень похожее на кнопки



магнитофона – те же *“Record”*, *“Play”*, *“Stop”*; при новом действии *«Запись»* нажимается автоматически) и делаем что угодно: например, сначала поднимаем контрастность и яркость (*Редактировать – Изображение – Яркость и контрастность*), потом убираем по полсантиметра по краям, потом сохраняем в *bitmap* (в 300 dpi), потом закрываем. Главное – не забыть нажать на кнопку *«Стоп»*. Всё, с остальными страницами можно не заморачиваться: нажимаем на *“Play”* – и готово. При этом стоит создать отдельные последовательности действий для текста и фототаблиц (их-то в битмап перегонять категорически не стоит). Да, если изначально файл сканировался в *bitmap* (иногда это проще, но получившиеся файлы обрабатывать сложнее, а для фототаблиц это опять же не подходит), то, например, для поворота изображения (*Файл – трансформация*, или просто *Ctrl+A* (выделить всё) + *Ctrl+T* (трансформировать)) придётся сначала перегонять его в *grayscale*, а потом, соответственно – обратно.

Многие хорошие вещи позволяет проделать над большим количеством TIFF-файлов программа [Scan Kromsator](#) - если собираетесь сканировать постранично много и хотите делать это качественно, используйте ее.

Также имеется более простая и удобная в работе программа, чем Scan\_Kromsator, особенно полезная при обработке сканов, изначально полученных в оттенках серого или книг, страницы которых были сфотографированы, а не отсканированы - [ScanTailor](#). Саму программу ScanTailor можно скачать [здесь](#), посмотреть краткое её описание [зут](#), а обучиться работе с программой можно [с помощью видеоурока](#), где создатель программы демонстрирует приёмы работы с ней. Программа автоматически устраняет перекося страниц, выделяет полезные области на страницах (текст/изображения), добавляет поля, быстро и качественно преобразует текстовую часть в черно-белую (если страницы сканировались в оттенках серого, например). В общем - очень полезная программа!

Ну вот, всё отсканировано, файлы успешно обработаны – и Вы с ужасом замечаете, что они занимают сотню-другую мегабайт. Не пугайтесь, при объединении в PDF сильнее всего сожмутся как раз самые здоровые файлы – фототаблицы, всякие цветные изображения, и итоговый файл будет намного, намного меньше (особенно если в финале его оптимизировать с помощью программы *Adobe Acrobat Professional*, например, так: *Document (File – в 6-7 версиях)– Reduce file size* – выставив совместимость с *Acrobat 5* (более древние варианты почти ни у кого не сохранились, в случае же чего они легко улучшаются до *Acrobat Reader 8*)

## Какой сканер использовать?

При покупке сканера первое, с чем следует определиться - с целью его применения. Если мы говорим о сканировании книг формата не крупнее А4, первое, что заслуживает упоминания - сканер должен быть планшетным (и никак иначе!). Протяжные и ручные агрегаты не подходят. Главными техническими характеристиками будут скорость сканирования и длительность полного цикла сканирования (время подготовки к сканированию + время сканирования + время обратного хода каретки).

**А) Какой сканер выбрать для сканирования толстых книг или большого количества статей?** Как ни странно, на этот вопрос существует однозначный ответ - *Plustek OpticBook*. Техника, почти идеально заточенная под сканирование книг с разрешением 300 dpi. Скорость сканирования страницы размером А4 в любом цветовом режиме (ч/б, оттенки серого, цветной) - 5-7 секунд для модели 3600, 2-3 секунды для модели 4600. Кроме того, близкие параметры имеют сканеры *Avision FB2280E (формат А4) / FB6280E ( формат А3)*. **Относительно недорогой сканер на основе цифровой камеры с большой скоростью сканирования - Sceye**

Прочие специализированные "книжные" сканеры находятся в совершенно другом ценовом классе и доступны скорее крупным учреждениям, чем частным лицам. В условиях редкости товара в магазинах бытовой и офисной техники и отсутствия конкуренции в данном сегменте рынка сканеров, диапазон цен на *Plustek* весьма обширен. Складывается ощущение, что цена на эти устройства обычно устанавливается продавцом "от балды". Адекватная цена для модели 3600 по состоянию на лето 2008 г. - около 7-8 тыс. руб. (можно легко разыскать и за 25-35 т.р.:)), модели 4600 - около 15-18 тыс. руб. Информацию о том, как и где приобрести технику *Plustek*, ищите на сайте компании <http://www.plustek.ru/> или на <http://price.ru>

Два мелких недостатка использования книжных сканеров *OpticBook*: 1) габаритность (толщина около 10-11 см) 2) зависимость от электропитания (как и для любого другого ССD-сканера, использующего флуоресцентные лампы). Это не позволяет носить его с собой постоянно, или, например, работать в библиотеке, где не всегда можно получить доступ к электросети. Кроме того, для использования в других повседневных целях (сканирования фотографий и фотопленок, например) эти модели не подходят из-за откровенно плохой цветопередачи.

**Б) Если нужен сканер для использования в разных целях, в том числе работы с цветными изображениями и фотопленками**, стоит приобрести ССD-сканер одной из привычных фирм - *Epson, Canon*. Новые и старые модели у большинства производителей по главной характеристике - скорости сканирования - различаются очень слабо, причем сравнение не всегда в пользу новых моделей, большинство наворотов которых заточено под главную массовую потребность - работу с фотографиями и фотопленкой, либо касаются интерфейса. Можно смело брать понравившуюся, пусть и самую дешевую, модель, только обращайтесь внимание на скорость сканирования. Причем, для ч/б, серого и цветного изображения она может сильно различаться, а может - и не различаться вовсе.

Габаритность - средняя. Питание от электросети.

**В) Если Вы не сканируете многостраничных книг, довольно редко работаете с цветными изображениями**, и сканирование сводится к необходимости иногда делать копии одной-нескольких статей в библиотеке, оптимальным решением будет CIS-сканер на светодиодах, который отличается малыми габаритами (толщина всего 2-4 см.), а электропитание обеспечивается по USB от компьютера. Идеальный вариант - ноутбук + CIS-сканер, например, из серии *Canon Lide*. Их минусы. Малая глубина резкости (на многих ССD-сканерах можно получать удобоваримые 3D-изображения мелких объектов, правда, со слегка искаженными пропорциями), по этой же причине при сканировании в ч/б книгу надо очень плотно прижимать к стеклу, иначе появляются черные полосы в тех местах, где это не достигается. Кроме того, CIS-сканеры в среднем работают медленнее, чем ССD, но в каждом конкретном случае все зависит от модели.

## Где достать необходимые программы?

Некоммерческие (т.е. бесплатные) – в Интернете. Прочие:

- приобрести у производителя
- купить на Горбушке или ее функциональном аналоге
- хорошенько поискать в Интернете. Многие программы по DJVU, а также кое-что по PDF можно найти на [сайте Виктора Иванова](#).
- найти в [файлообменных сетях](#) (например, e2dk) через опцию «Поиск» [программы-клиента](#).

## Ничего не получается, есть важные дополнения или остались вопросы?

Загляните на [форум сайта JURASSIC.RU](#) (→Общая информация, вопросы и ответы → Все об электронной литературе).



Ваши замечания и предложения по оформлению и содержанию сайта пишите [зюда](#)

Последнее обновление 03.10.2010

Текст страницы: [Ипполитов А.П.](#), [Рогов М.А.](#), [Гужов А.В.](#)