# Получение текстового слоя при помощи FineReader 11 и DjvuOCR



**Описан метод получения корректного текстового слоя без двойных пробелов в DJVU- книге при помощи FineReader 11.0.0.583 и DjvuOCR 2.4 beta RC4 mod NBell.**

## *Введение*

Почему FineReader 11 - ведь есть рабочая связка Finereader 8 + DjvuOCR?

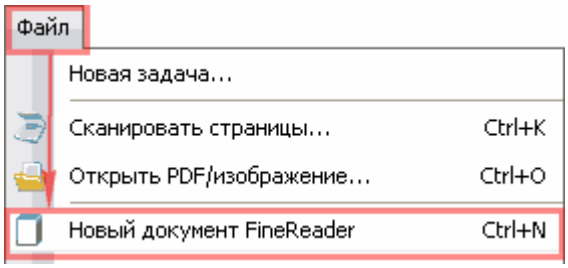
Ответ прост – DjvuOCR накладывает жесткие ограничения на проект Finereader 8 - редактирование зон при распознавании, исправление и добавление текста часто приводят к невозможности извлечения текста.

Finereader 11 сам умеет создавать текстовый слой, не накладывая практически никаких ограничений на распознавание и правки текста.

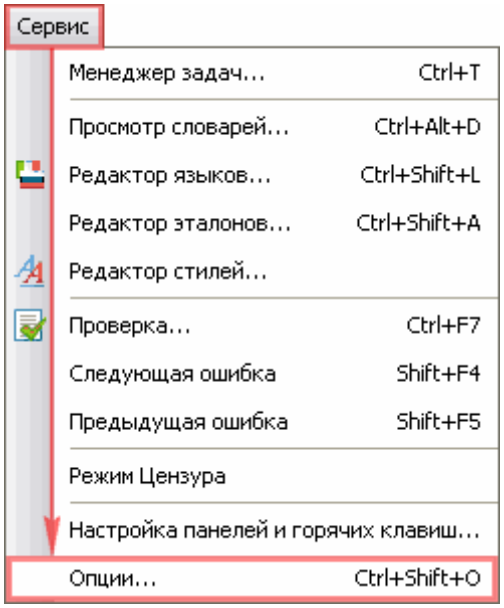
Поэтому при необходимости качественно править текст в OCR-слое нижеприведенная методика единственно удобная и приемлемая (на текущий момент – может быть ABBYY сделает полноценный DjVu-кодер в Finereader 12 с поддержкой раздельного кодирования текста и картинок).

### *Этап 1. Распознавание в Finereader*

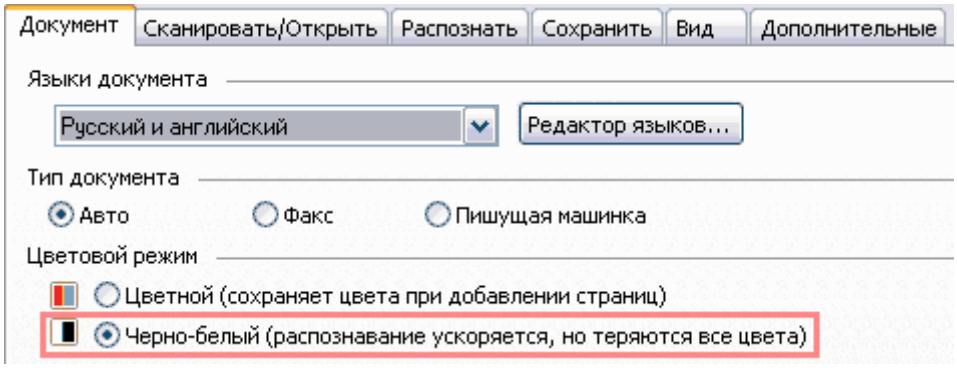
Создаем новый документ FineReader:



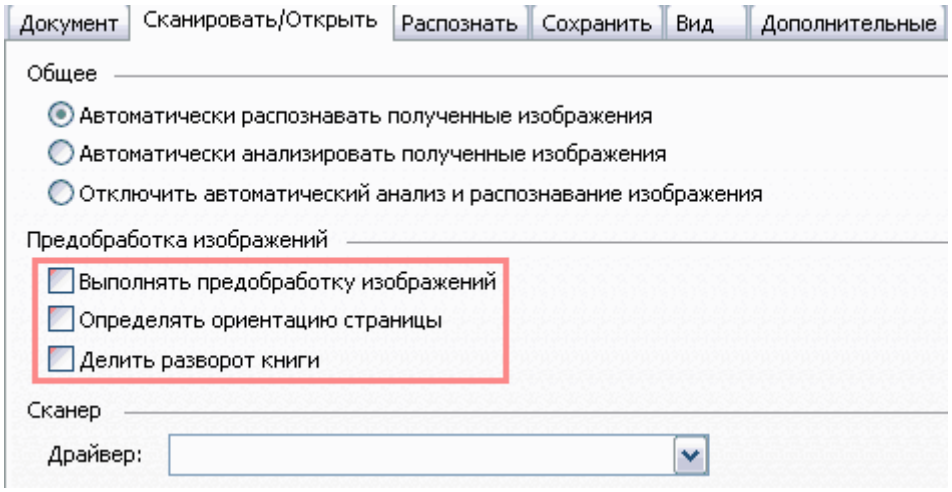

Устанавливаем опции согласно картинке:



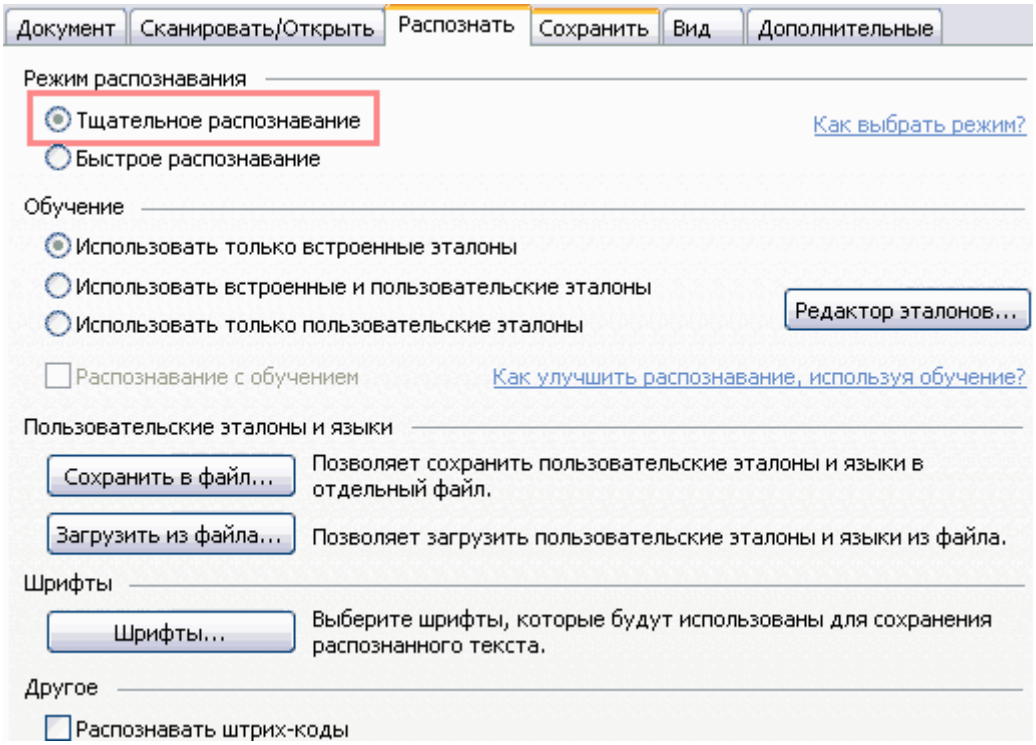
Языки для документа выберите по вкусу. "Русский и английский" я выбрал, так как текст технический и содержит много латинских обозначений. "Черно-белый" – мне не нужен полученный djvu как книга, только как источник текста.




Обратите внимание на выделенную красным группу опций – я распознавал уже выровненные сканы, поэтому искажение изображения мне не нужно. Что лучше для Вас – определите опытным путем.

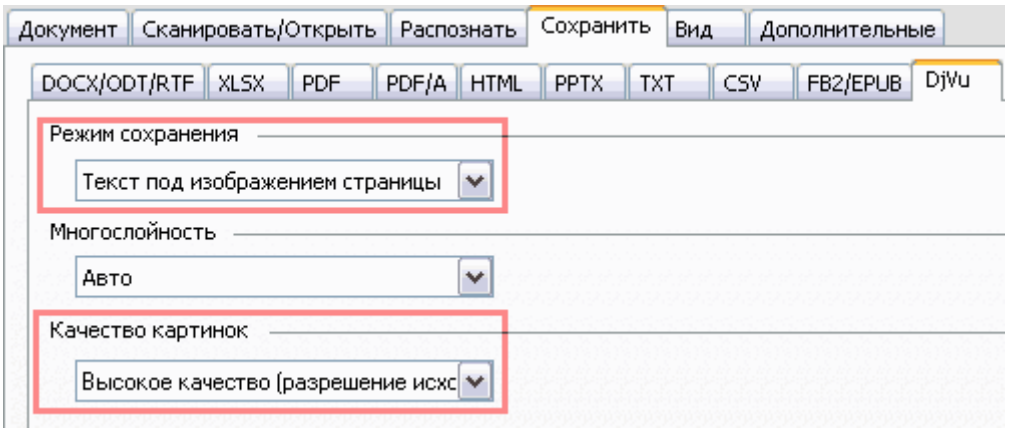


На этой вкладке я выбрал "Тщательное распознавание" – результат мне более важен, чем время.

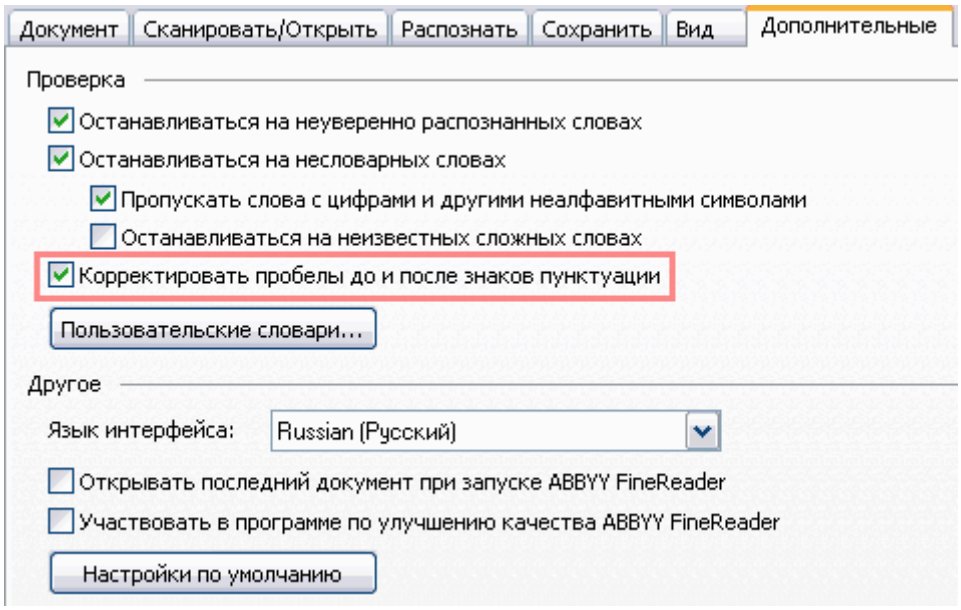



Далее импортируйте в проект исходные файлы, из которых сделан Ваш djvu (можно даже импортировать сам djvu).

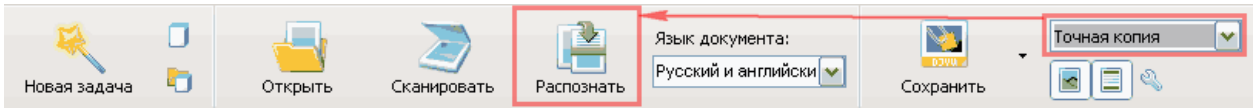
Настройки DjVu важны. Установите согласно картинке:



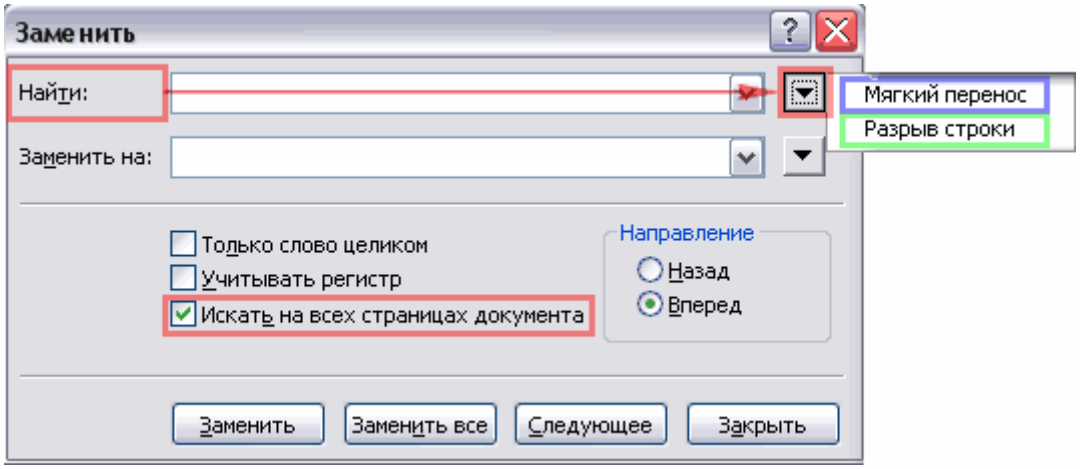
Корректировка лишних пробелов явно не лишняя:




Далее устанавливаем опцию "Точная копия", распознаем:

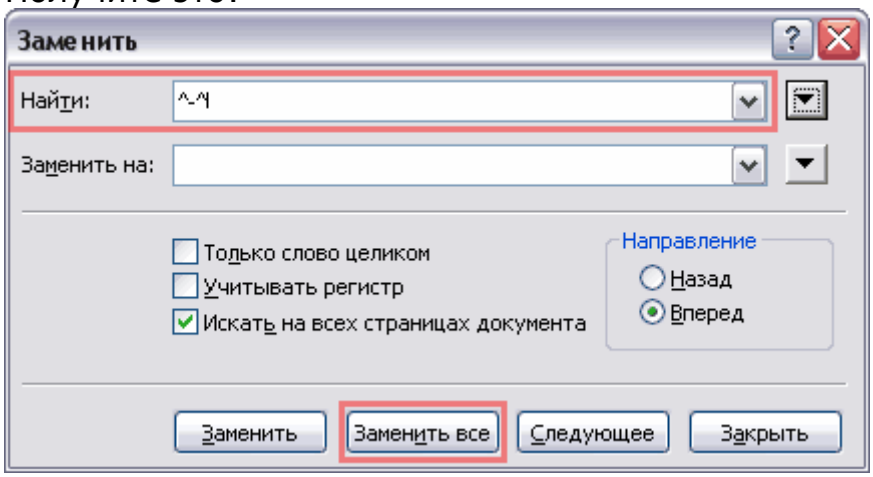



Примечание: удаления мягких переносов по этому способу не будет – DjvuOCR делает это при импорте из проекта Finereader версий 7-8, 9.0.0.724). Поскольку DjvuOCR не умеет работать с документами Finereader, то и переносы удалять не будет! Переносы придется удалять вручную в Finereader вот так:
Вызовите Ctrl+H окно "Заменить"




Установите опцию "Искать на всех страницах документа".
В выпадающем меню поля "Найти" выберите сначала "Мягкий перенос", затем "Разрыв строки".

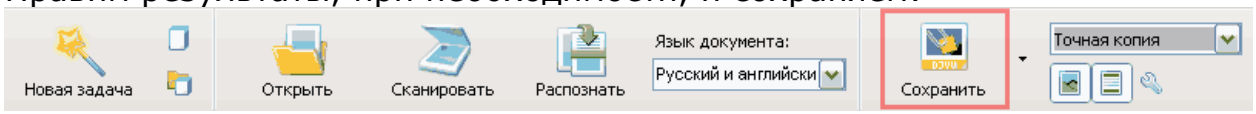
Получите это:




Нажмите "Заменить все".

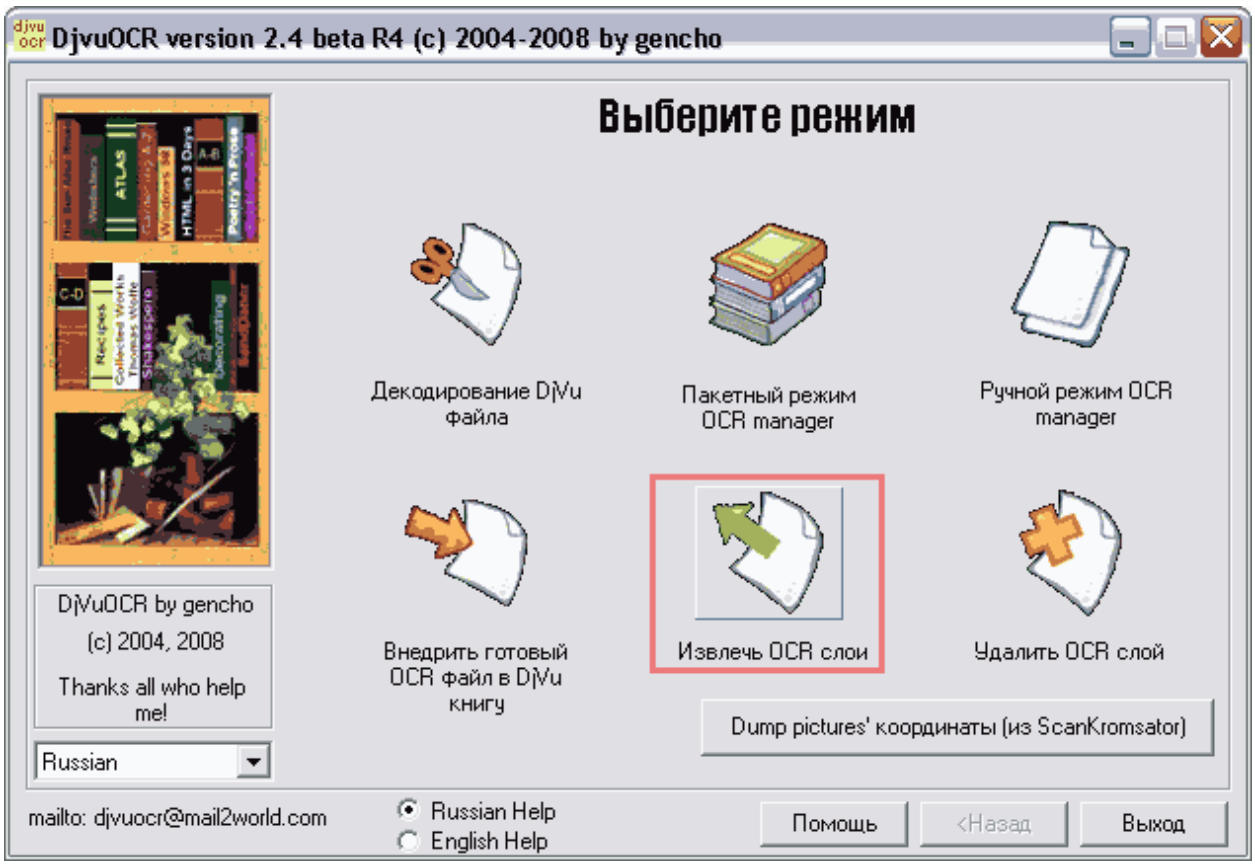
Можно также удалить и тире+"разрыв строки" – чаще всего это нераспознанный перенос. Но будьте внимательны – надо следить за тем, чтобы не были удалены настоящие тире. Поэтому рекомендую проводить такую замену в ручном режиме – нажимая кнопку "Следующие" и лишь затем – "Заменить".

Правим результаты, при необходимости, и сохраняем:

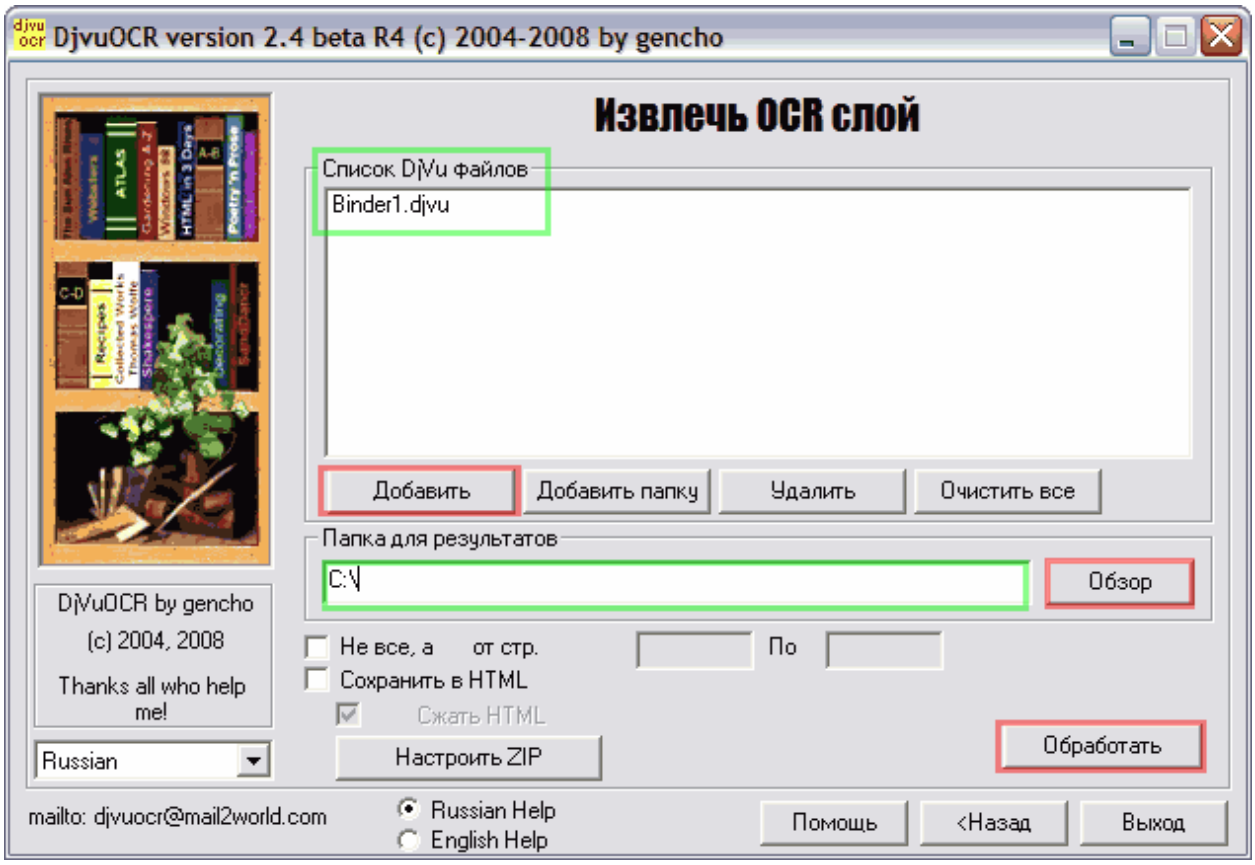


### *Этап 2. Перенос текста с помощью DjvuOCR*

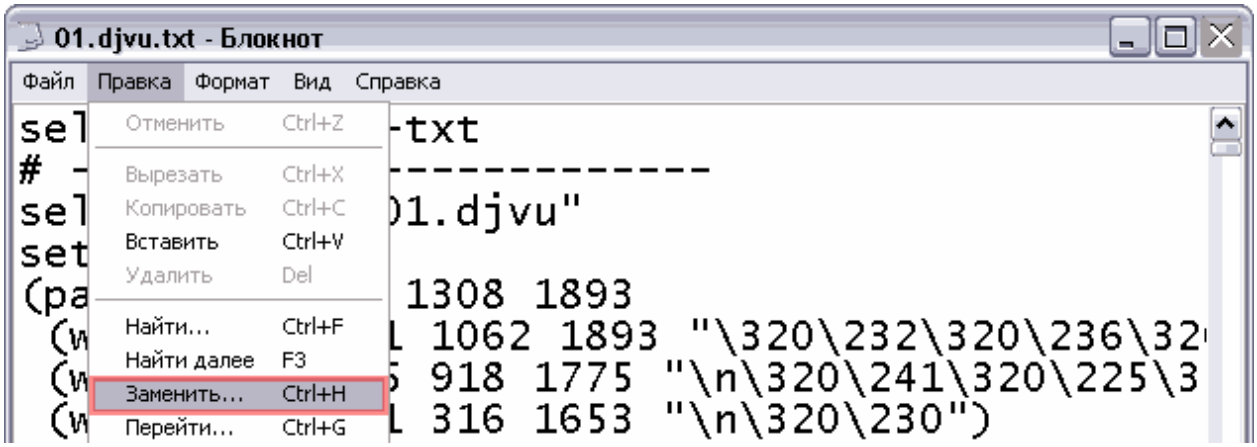
Запускаем программу DjvuOCR и нажимаем "Извлечь OCR слой":



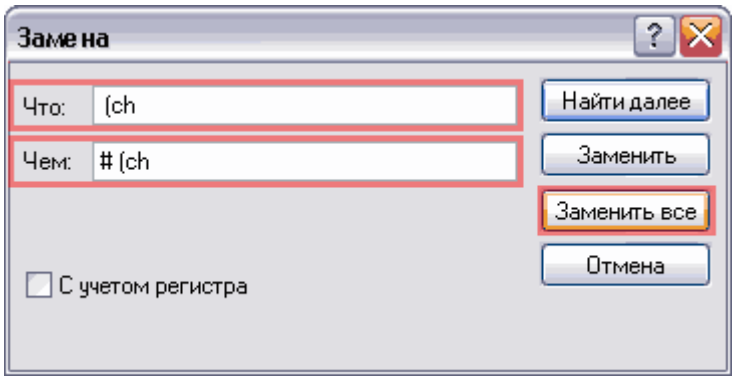
Кнопкой "Добавить" в стандартном диалоговом окне выбираем djvu, произведенный Finereader, и, выбрав папку назначения кнопкой "Обзор", нажимаем "Обработать":



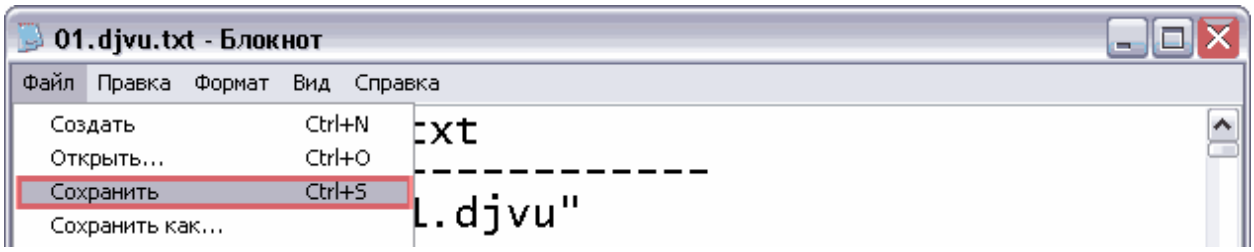

Открываем полученный txt-файл (это будет имяфайла.djvu.txt) в текстовом редакторе, поддерживающим UTf-8, например "Блокнот" и выбираем команду "Заменить" Ctrl+H:



В окне "Замена" в поле "Что" набираем (без кавычек!) " (ch", в поле "Чем" - "# (ch" и нажимаем кнопку "Заменить все"




Сохраняем файл командой Ctrl+S:




Затем внедряем полученный качественный текстовый слой в нужный Вам djvu. Кнопкой "<Назад" вернемся в главное окно и нажмем "Внедрить готовый OCR файл в DjVu книгу":

Затем в группе "OCR файл создан из проекта Finereader" кнопкой "Обзор выбираем Ваш файл с подготовленным текстовым слоем, в группе "В какой DjVu файл добавлять" кнопкой "Обзор" выбираем djvu книгу, для которой Вы готовили текстовый слой и нажимаем "Обработка":


Вот и все!

### *Послесловие*

Все так же как и после удаления переносов DjvuOCR при импорте текста из проекта Finereader 7-9 – весь текст слова ДО мягкого переноса переносится в его окончание – ту часть, что ПОСЛЕ переноса, которая на другой строке.
Нам бы радоваться – поиск по тексту полноценный и мягкие переносы убраны, но…
Полученный текстовый слой несколько хуже, чем при использовании Finereader 7-9 и обработки DjvuOCR.
Всплывет не очень существенная проблема выделения текста для копирования:

| В djvu после импорта текста из проекта Finereader 7-9 и обработки DjvuOCR: |
|---|

При выделении части слова ДО переноса не копируется ничего (в текстовом слое вместо начала слова до переноса пробел).
Выделение ПЕРЕНЕСЕННОЙ части слова и копирование – дает все слово.

| WinDjView и djvu с перенесенным текстовым слоем: |
|---|

Текстовый слой от Finereader 11 он плохо понимает – выделить слово - проблема… С трудом выделяются не более 2 страниц текста…

| Caminova/Lizardtech DjVu browser plug-in и djvu с перенесенным текстовым слоем: |
|---|

Та же проблема, что и с WinDjView.

| DjView DjvuLibre и djvu с перенесенным текстовым слоем: |
|---|

Выделяет текст нормально.

| В djvu от Finereader 11 после удаления переносов по моей методике: |
|---|

Начало слова не выделяется, только если выделить слово до и слово после.
Выделение и копирование окончания слова – дает все слово целиком.
Если Вас устраивает качество кодирования Finereader 11 – пользуйтесь djvu, кодированным Finereader 11. С ним меньше проблем при выделении текста…

| И еще неприятность – DjvuOCR не может сгенерировать HTML из djvu с текстовым слоем в стиле Finereader 11. |
|---|