Научно-исследовательское и просветительное учреждение культуры
Национальный Полоцкий историко-культурный музей-заповедник

Научно-информационный отдел

Методические рекомендации
Выпуск 6

РЕКОМЕНДАЦИИ ПО СКАНИРОВАНИЮ И РАСПОЗНАВАНИЮ ТЕКСТА

Полоцк 2013

Оглавление

Оцифровка текста — это процесс перевода текста с бумажного носителя (обычно через сканер) в электронный (цифровой) вид.

Методика оцифровки

Процесс оцифровки включает **два этапа**:

I. Получение копий страниц в виде графических (растровых) изображений, осуществляемое путём сканирования с последующей обработкой и сохранением в ***формате графических файлов — .tif***. В этом случае полностью исключаются какие-либо ошибки, однако невозможен поиск или извлечение фрагментов текста, например, для цитирования.

II. Распознавание текста при помощи специальной программы (технология «оптического распознавания символов»[1]) с последующим сохранением в одном из ***текстовых форматов — .rtf***, ***.docx***, ***.odt***.

Существуют некоторые различия при сканировании одной страницы (🗋) и многостраничного документа (🗐).

I этап. Сканирование текстовых страниц

Части страницы, где находится текст, должны быть полностью прижаты к стеклу, иначе возникает характерное затемнение (в книге — в области корешка) из-за наклонного падения света лампы в сканере.

Необходимо определить оптимальную рамку и параметры сканирования.

1) Установка области сканирования.

🗋 Выделите рамкой в окне просмотра программы сканирования только область текста (если нет необходимости сохранить точное форматирование страницы).

1 Оптическое распознавание символов (англ. *optical character recognition, далее — OCR*) — перевод графических изображений текста в последовательность кодов, использующихся в ***текстовом редакторе***. Оптическое распознавание текста позволяет ***редактировать текст***, осуществлять поиск, хранить его в более компактной форме, а также применять к тексту электронный перевод.

🗍 Выделите всю страницу при первом сканировании. Не рекомендуется изменять размеры выделенной области при последовательном сканировании нескольких листов — программа OCR в этом случае выдаст для каждого листа свой размер бумаги, что затруднит печать распознанных документов. Область сканирования выставляется с небольшим запасом относительно формата страницы, чтобы не особенно заботиться о точном выравнивании книги на стекле.

2) Выбор настроек сканера:

Для оптического распознавания ***нормального типографского текста***: разрешение 300*300 dpi, с 8-bit серого (с 24-bit цвета, если есть цветные иллюстрации), с сильным повышением чёткости, с режимом оцифровки «Документ», масштаб 100%.

В окне просмотра текст должен быть «читаем», то есть текст имеет ровный контур, нет затемненных областей, искажений изображений и фона от бумаги. Для устранения этих погрешностей используйте регуляторы контраста и яркости.

3) Сканирование и сохранение изображений.

Порядок выполнения стандартный *(см. «Методические рекомендации. Выпуск 4»)*. Создайте папку для сохранения отсканированных изображений страниц, назначьте имя и порядковый номер для первого файла (*например:* page0001.tif).

🗍 В конце просмотрите все страницы и исправьте недостатки сканирования. Например, иногда книга неровно легла на стекло и часть текста на какой-либо странице не отсканировалась или были вовсе пропущены некоторые страницы.

II этап. Распознавание текста

1) Загрузка отсканированных изображений в программу OCR.

Запустите программу оптического распознавания теста. Выберите в меню «Файл» пункт «Открыть» или соответствующую кнопку на панели управления.

Найдите и выделите все заранее отсканированные изображения.

Нажмите кнопку OK. Изображения загрузятся в программу распознавания.

Выделите страницу, которую необходимо распознать. Она загрузится в окно просмотра.

После этого в некоторых программах можно немного подредактировать (с помощью соответствующих инструментов) изображение: повернуть его, убрать «шум», «мусор» и т.п.

2) Разметка.

Разметка нужна для правильного выделения на рисунке областей с текстом, таблицей, рисунком, а также областей, которые не нужно отображать. Возможно выполнить разметку ***автоматически***, а после редактировать её вручную. Не пренебрегайте редактированием разметки — при сложной вёрстке (текст в несколько колонок или блоками) неправильная разметка может сделать текст нераспознаваемым.

Среди параметров редактирования есть такие, как:

- создание новых блоков;
- удаление существующих блоков;
- изменение ***типов*** существующих блоков;
- изменение размеров блоков;
- добавление пространства к блоку;
- удаление пространства из блока.

3) Выбор языка распознавания.

Для правильного распознавания символов программой OCR необходимо установить язык (или несколько языков) распознавае-

мого текста в соответствующем меню.

4) Распознавание текста.

🗋 Запустите процесс распознавания. Распознавание обычно ведётся в ***автоматическом режиме***.

🗍 Перед началом распознавания вернитесь к первой странице документа. Повторите эти операции для других листов. При переходе с одного листа на другой программа может запросить разрешение на запись сделанных изменений.

5) **Проверка текста** *(рекомендуется, если в программе есть соответствующая функция)*.

Проверьте орфографию и оформление распознанного текста. Несмотря на трудоёмкость процесса, это позволит значительно сократить время на дальнейшую обработку в текстовом редакторе. Для правки текста необходимо запустить проверку орфографии кнопкой «Проверить» (далее следуйте инструкциям программы).

6) Сохранение.

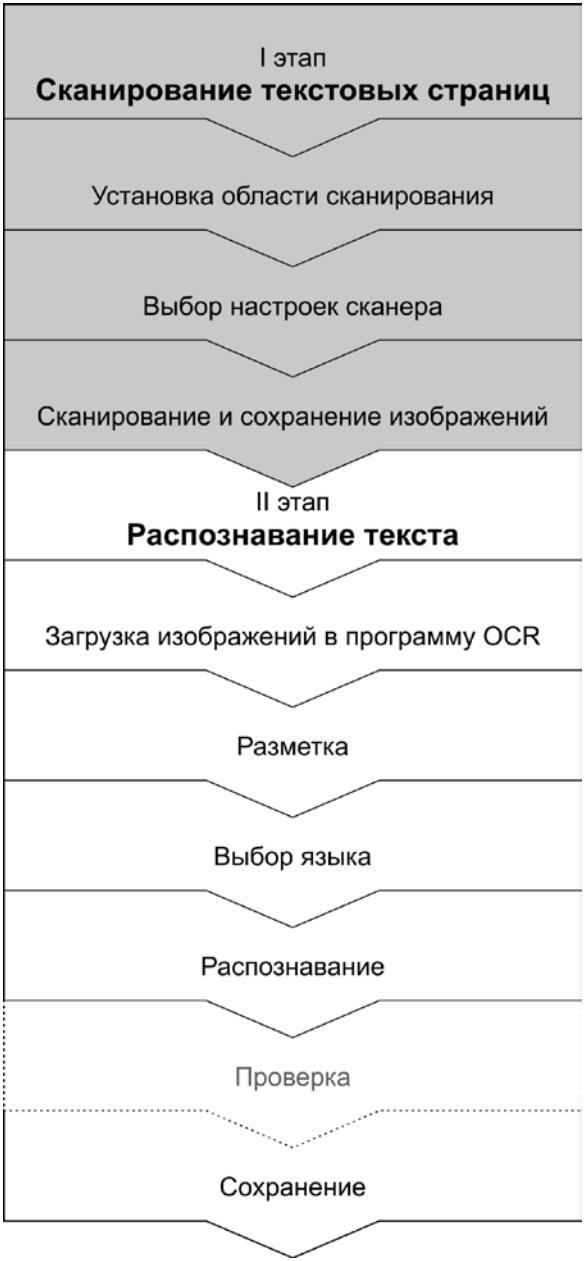
Экспорт текста из пакета происходит после выбора кнопки «Сохранить». При этом запустится мастер сохранения, который запросит, куда надо экспортировать текст:

- сохранить ли его в файле (форматы .rtf, .docx, .odt, .txt и др.);

🗍 можно сохранить либо все листы в один файл, либо сохранить каждый лист в отдельном фале;

- передать текст в другую программу (текстовый редактор, программу-переводчик и т.п.);

Алгоритм процесса сканирования и распознавания текста



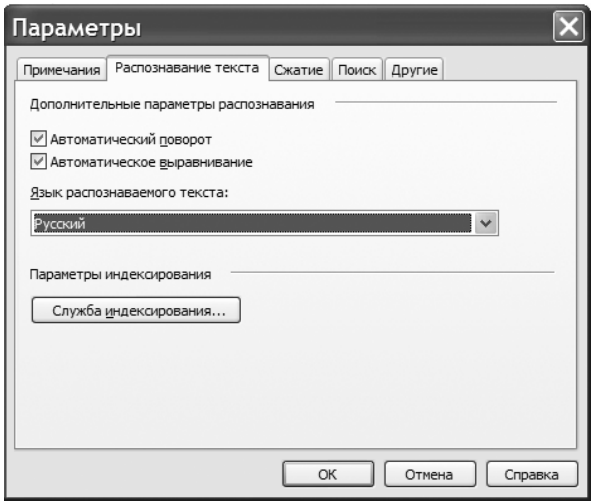
Приложение 1. **Распознавание текста при помощи программы Microsoft Office Document Imaging**

Данная программа входит в пакет «Microsoft Office». Набор функций в ней ограничен. Например, возможно распознавание только для одного языка и для текста простой компоновки (в одну колонку).

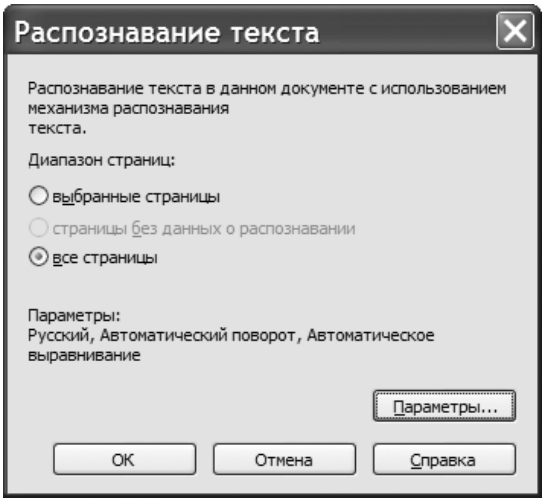
1) Запустите программу: *Пуск → Все программы → Microsoft Office → Средства Microsoft Office → Microsoft Office Document Imaging.*

2) В меню «Файл» выберите пункт «Открыть…» и в диалоговом окне укажите файлы отсканированных страниц.

3) В меню «Сервис» кликните пункт «Параметры…» и в диалоговом окне на вкладке «Распознавание текста» выберите (из предложенных) язык распознаваемого текста.

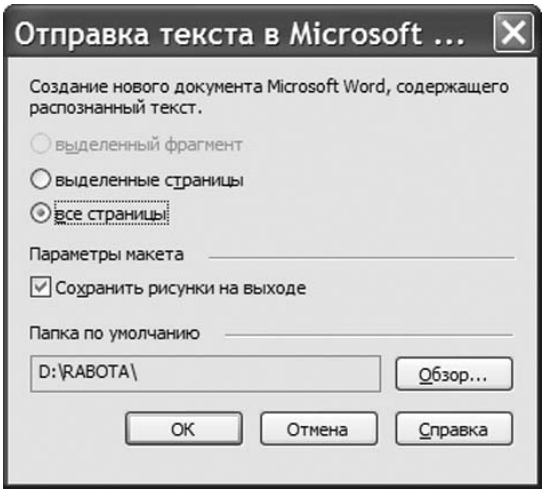


4) Выделите первую страницу на боковой панели. В меню «Сервис» выберите «Распознать текст…» и в диалоговом окне отметьте пункт «все страницы».



Кнопка «ОК» запустит автоматический процесс распознавания.

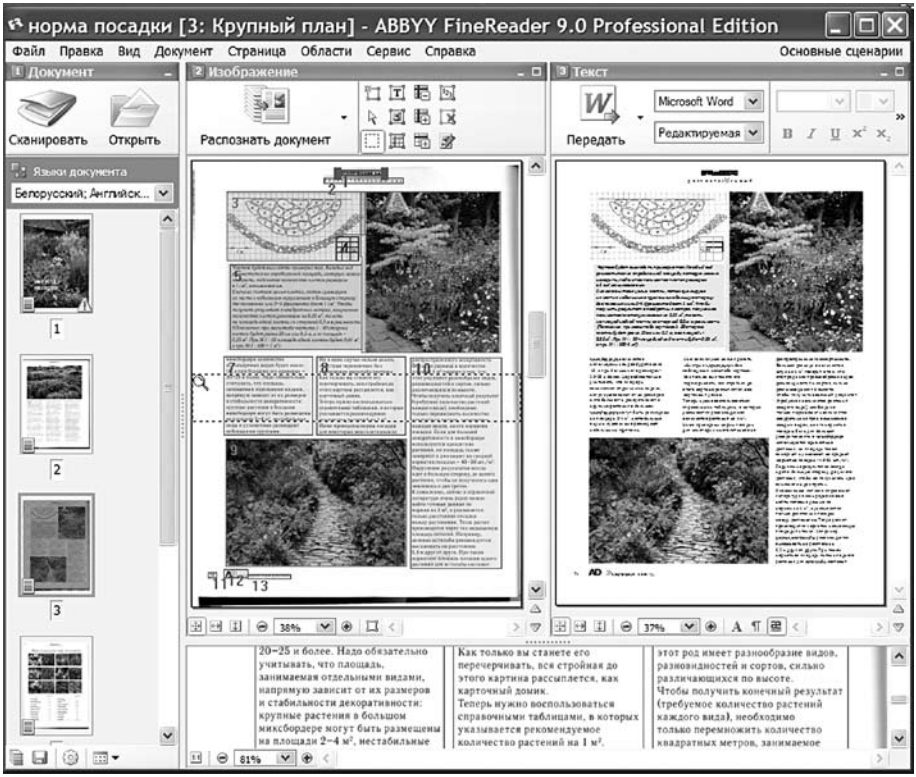
5) В меню «Сервис» выберите «Отправить текст в Microsoft Word…». В диалоговом окне отметьте пункт «все страницы» и, если необходимо, «Сохранить рисунки на выходе». Укажите папку для сохранения файла.



Кнопка «ОК» запустит процесс экспорта в текстовый редактор, где вы сможете внести необходимую правку и отформатировать материал. Сохраните файл.

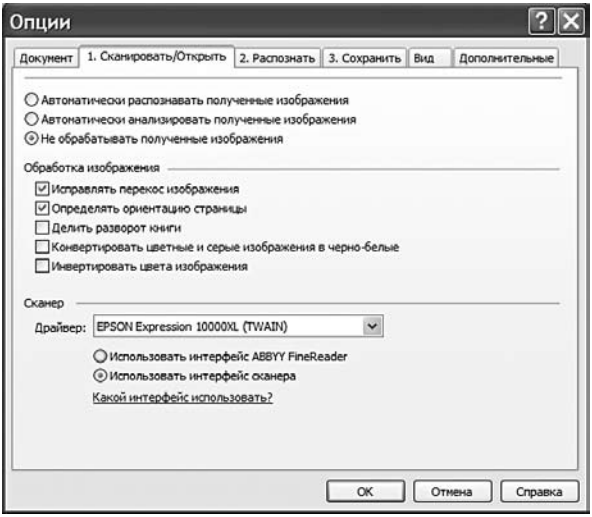
Приложение 2. **Распознавание текста при помощи программы ABBYY FineReader 9.**

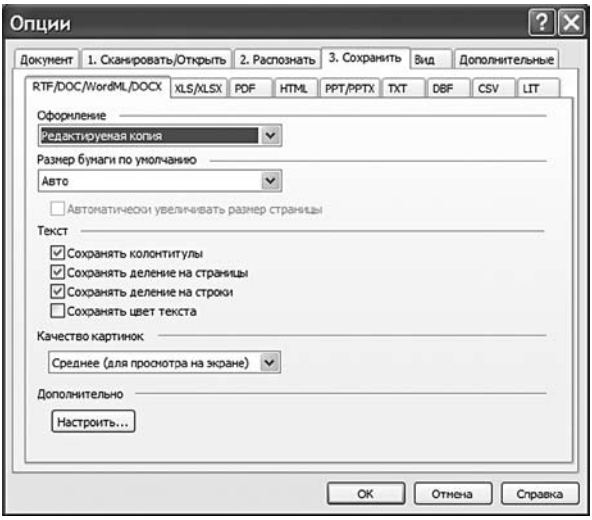
1) Запустите программу.



1) В меню «Файл» или на панели инструментов выберите «Открыть…» и в диалоговом окне укажите файлы отсканированных страниц. Нажмите «ОК». Страницы загрузятся в программу (эскизы в окне «Документ»).

2) В меню «Сервис» выберите пункт «Опции». В диалоговом окне внесите следующие изменения на вкладках «Сканировать/Открыть» и «Сохранить»:





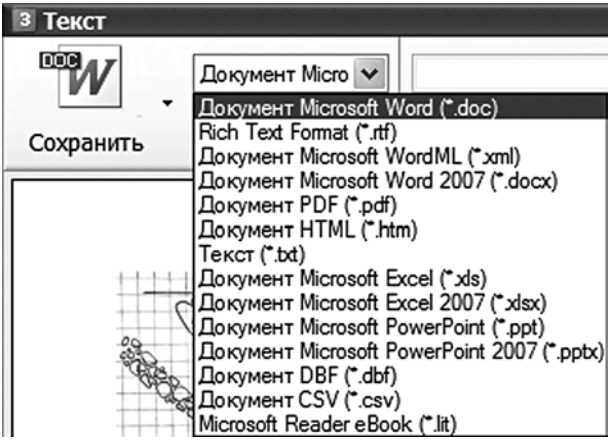
3) В окне «Документ» или через *Сервис → Опции → Документ* выберите один или несколько языков распознаваемого текста.

4) Выделите первую страницу и выполните анализ документа (*Документ → Анализ документа* или кнопка на панели иструментов). В окне «Изображение» можно отредактировать блоки «Текст», «Таблица», «Рисунок» и др. для последующего правильного распознавания.

5) Нажмите на кнопку «Распознать документ». Распознавание проходит в автоматическом режиме, неуверенно распознанные символы выделяются синим цветом в окне «Текст».

6) Программа позволяет выполнить проверку текста (*Сервис → Проверка*).

7) В окне «Текст» на панели инструментов в выпадающем списке выберите формат файла для сохранения.



8) Нажмите на стрелку «▾», находящуюся справа от кнопки «Сохранить», далее в меню выберите способ сохранения распознанного текста. Нажмите на «Сохранить».

Рекомендации по сканированию изображений
Методические рекомендации
Выпуск 6

Составитель
Урбан Светлана Геннадьевна

Корректор *В.Е. Ошуева*

Формат 60x84 1/16. Бумага Navigator. Гарнитура «Times»
Печать лазерная. Тираж 50 экз.

Напечатано на оборудовании
научно-информационного отдела НПИКМЗ
211400, г. Полоцк, ул. Нижне-Покровская, 22