# Системы оптического распознавания документов

# Необходимость в системах распознавания символов

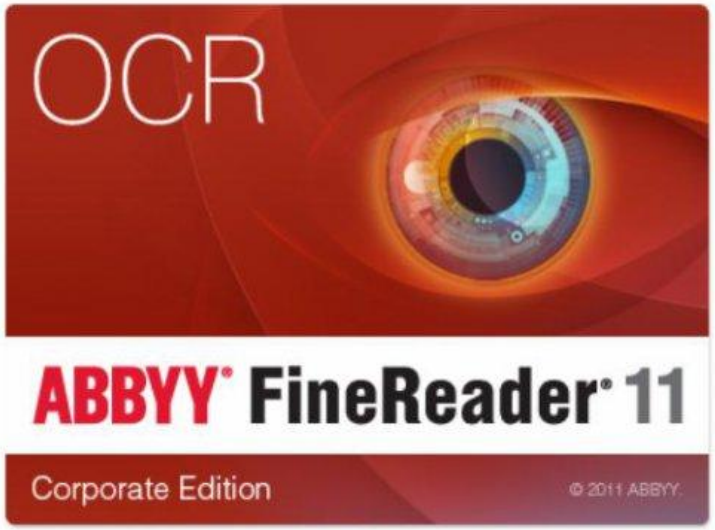
- С помощью сканера достаточно просто получить изображение страницы текста в графическом файле. Однако работать с таким текстом невозможно: как любое сканированное изображение, страница с текстом представляет собой графический файл - обычную картинку. Текст можно будет читать и распечатывать, но нельзя будет его редактировать и форматировать. Для получения документа в формате текстового файла необходимо провести распознавание текста, то есть преобразовать элементы графического изображения в последовательности текстовых символов.

# Программы распознавания текста

Преобразованием графического изображения в текст занимаются специальные программы распознавания текста (Optical Character Recognition - **OCR**).

Наиболее распространенные системы оптического распознавания символов:

- ABBYY FineReader
- CuneiForm от Cognitive



# Получение электронного документа

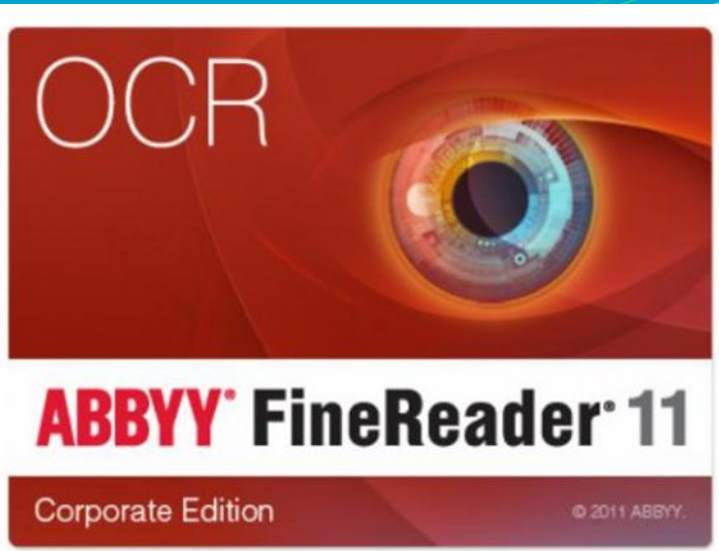

1. Отсканировать изображение (с помощью ПО сканера);
2. Распознать структуру размещения текста на странице: выделить колонки, таблицы, изображения и т.д.
3. Выделенные текстовые фрагменты графического изображения страницы необходимо преобразовать в текст;
4. Проверка орфографии (если необходимо);
5. Сохранение в файл или передача текста в другое приложение, например в Word.

# Методы распознавания символов

- Если исходный документ имеет типографское качество то задача распознавания решается **методом сравнения с растровым шаблоном**.

- При распознавании документов с низким качеством печати используется метод распознавания символов **по наличию** в них **определенных структурных элементов** (отрезков, колец, дуг и др.).

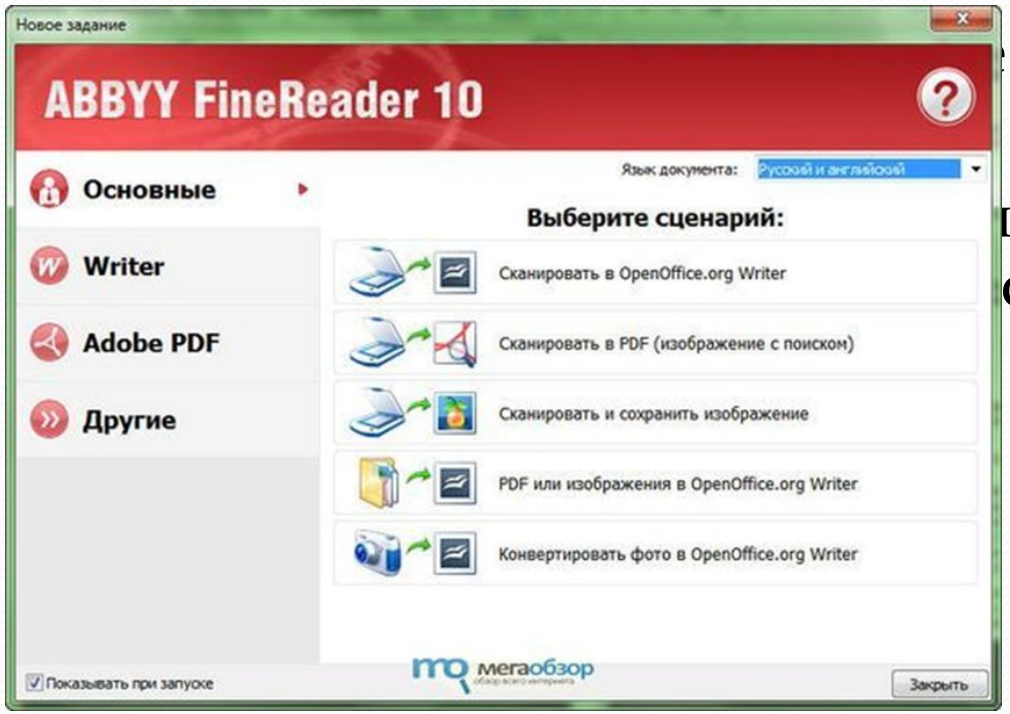
# ABBYY FineReader




FineReader - омнифонтовая система оптического распознавания текстов. Это означает, что она позволяет распознавать тексты, набранные практически любыми шрифтами, без предварительного обучения. Особенностью программы FineReader является высокая точность распознавания и малая чувствительность к дефектам печати.

# Оптимальное разрешение при сканировании

- обычный текст - 300 dpi
- мелкий шрифт (9 и менее пунктов)- 400-600 dpi



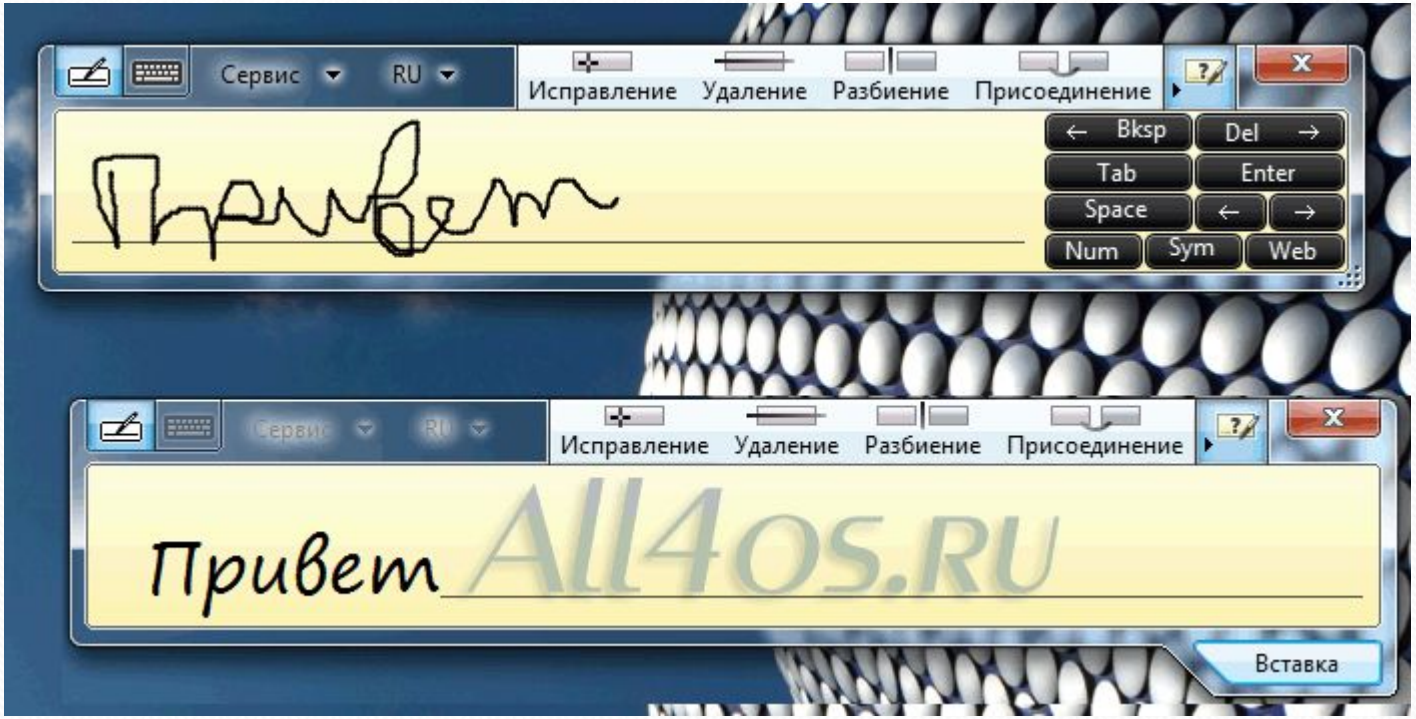

подбор яркости.

итов (картинки, цвет букв и

ой тип изображения.

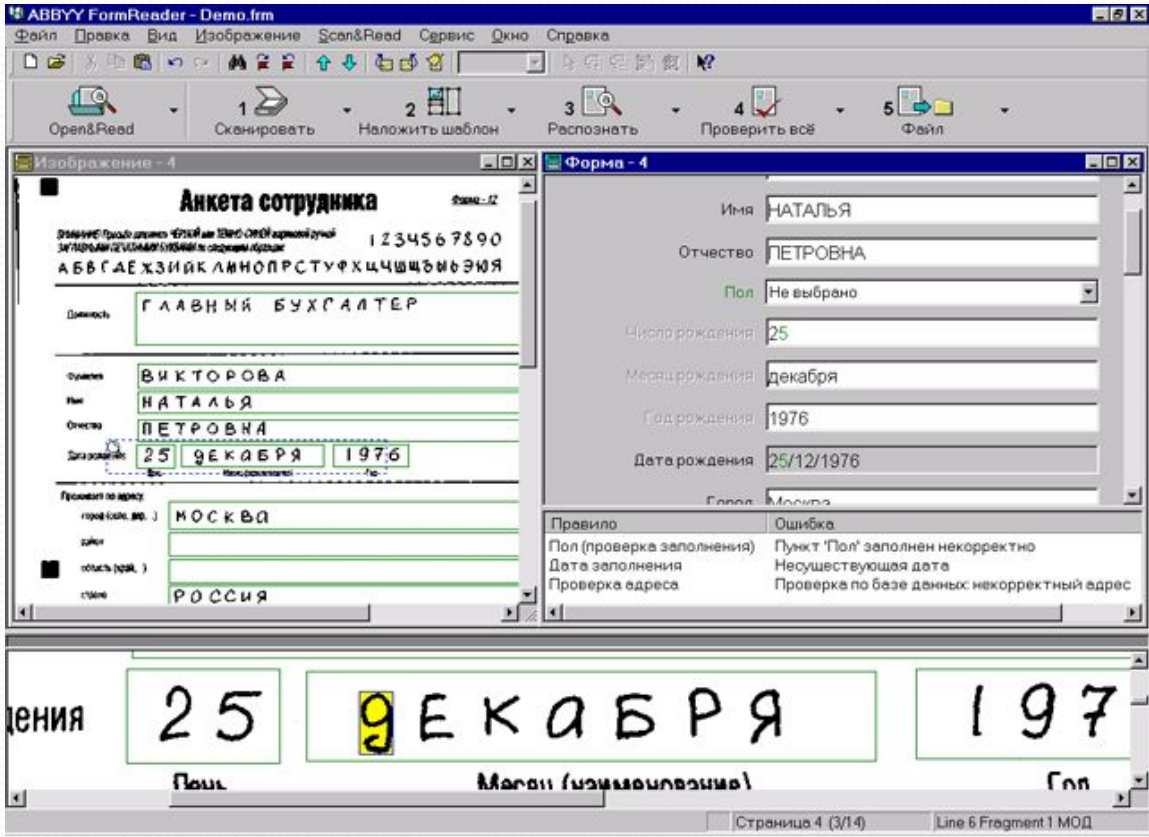
# Системы распознавания рукописного текста

- преобразуют текст, созданный на экране карманного компьютера специальной ручкой, в текстовый компьютерный документ.



# Системы оптического распознавания форм

- При заполнении документов большим количеством людей (например, при сдаче (ЕГЭ)) используются бланки с пустыми полями. Данные вводятся в поля печатными буквами от руки. Затем эти данные распознаются с помощью систем оптического распознавания форм и вносятся в компьютерные



# Вопросы:

- Зачем нужны программы распознавания текста?
- Как происходит распознавание текста?
- Какие программы распознания текста вы знаете? Какими пользовались?
- Какое разрешение  является оптимальным для сканирования текста, изображений?

# Домашнее задание:

- §2.8, вопросы.