

# Системы оптического распознавания документов





# Системы оптического распознавания символов

При создании электронных библиотек и архивов путем перевода книг и документов в цифровой компьютерный формат, при переходе предприятий от бумажного к электронному документообороту, при необходимости отредактировать полученный по факсу документ используются системы оптического распознавания символов.





# Оптическое распознавание СИМВОЛОВ

## Оптическое распознавание символов

(англ. optical character recognition, OCR) — механический или электронный перевод изображений рукописного, машинописного или печатного текста в последовательность кодов, использующихся для представления в текстовом редакторе.

С помощью сканера несложно получить изображение страницы текста в графическом файле.

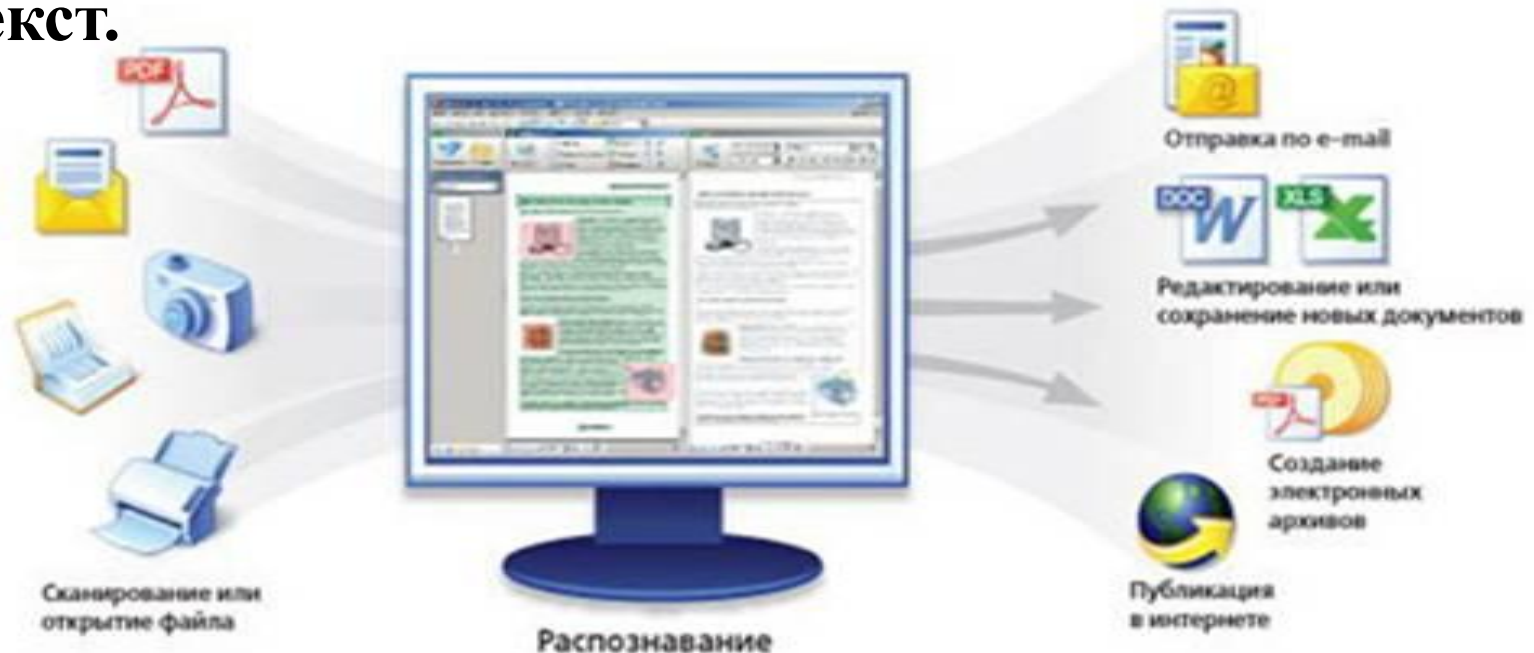


Однако для получения документа в формате текстового файла необходимо провести **распознавание текста**, т. е. преобразовать элементы графического изображения в последовательности текстовых символов.





- Сначала необходимо **распознать структуру** размещения текста на странице: выделить колонки, таблицы, изображения и т. д.
- Далее выделенные текстовые фрагменты графического изображения страницы необходимо **преобразовать в текст**.





## *Хорошее качество текста*

# **Растровый метод распознавания текста**

Если исходный документ имеет типографское качество (достаточно крупный шрифт, отсутствие плохо напечатанных символов или исправлений), то задача распознавания решается методом сравнения с растровым шаблоном.





## *Хорошее качество текста*

### **Растровый метод распознавания текста**

- Сначала растровое изображение страницы разделяется на изображения отдельных СИМВОЛОВ.
- Затем каждый из них последовательно накладывается на шаблоны символов, имеющихся в памяти системы, и выбирается шаблон с наименьшим количеством точек, отличных от входного изображения.

**А В В Ф Я**



## *Хорошее качество текста*

# Растровый метод распознавания текста

- Растровое изображение каждого символа последовательно накладывается на растровые шаблоны символов, хранящиеся в памяти системы оптического распознавания. Результатом распознавания является символ, шаблон которого в наибольшей степени совпадает с изображением



Например, распознаваемый символ "Б" накладывается на растровые шаблоны символов (А, Б, В и т. д.)



## *Плохое качество текста*



### **Структурный метод распознавания**

- При распознавании документов с **низким** качеством печати (машинописный текст, факс и т.д.) используется **метод распознавания структурных элементов** (отрезков, колец, дуг и др.) символов. В искаженном символьном изображении выделяются характерные детали и сравниваются со **структурными шаблонами** символов.

**И. Н**

- Любой символ можно описать через набор параметров, определяющих взаимное расположение его элементов. Например, буква «Н» и буква «И» состоят из трех отрезков, два из которых расположены параллельно друг другу, а третий соединяет эти отрезки. Различие между буквами в величине углов, которые составляет третий отрезок с двумя другими.

## *Плохое качество текста*



### **Структурный метод распознавания**

При распознавании структурным методом в искаженном символьном изображении выделяются характерные детали и сравниваются со структурными шаблонами символов.

В результате выбирается тот символ, для которого совокупность всех структурных элементов и их расположение больше всего соответствуют распознаваемому символу.



Например, распознаваемый символ "Б" накладывается на векторные шаблоны символов (А, Б, В и т. д.)



## Системы оптического распознавания форм

При проведении **Единого государственного экзамена**, при заполнении налоговых деклараций и т. д. используются различного вида бланки с полями. Рукописные тексты (данные вводятся в поля печатными буквами от руки) распознаются с помощью **систем оптического распознавания форм** и вносятся в компьютерные базы данных.

Сложность состоит в том, что необходимо распознавать символы, написанные от руки, а они довольно сильно различаются у разных людей. Кроме того, система должна определить, к какому полю относится распознаваемый текст.



# Системы оптического распознавания форм

Единый государственный экзамен - 2005  
**Бланк регистрации**

Заполнить геленой или капиллярной ручкой ЧЕРНЫМИ чернилами ЗАГЛАВНЫМИ ПЕЧАТНЫМИ БУКВАМИ по следующим образцам:  
А В В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Э Ю Я 1 2 3 4 5 6 7 8 9 0 X V I I L

**ВНИМАНИЕ!** Данный бланк использовать только совместно с двумя другими бланками из данного пакета

Сведения об участнике единого государственного экзамена

Фамилия \_\_\_\_\_  
Имя \_\_\_\_\_  
Отчество \_\_\_\_\_

Документ \_\_\_\_\_ Сервис \_\_\_\_\_ Номер \_\_\_\_\_ Пол:  Ж.  М.

Резерв - 1 \_\_\_\_\_ Резерв - 2 \_\_\_\_\_ Резерв - 3 \_\_\_\_\_ Факт выхода из аудитории во время экзамена

**ЗАМЕЧАНИЯ участника ЕГЭ по процедуре проведения ЕГЭ**

Заполнение НЕ ОБЯЗАТЕЛЬНО.

Отметьте  замечания по проведению экзамена:

<input type="checkbox"/> Организация доставки участника в пункт проведения ЕГЭ при самостоятельном времени в пути более 1 часа	<input type="checkbox"/> Присутствие в аудитории преподавателя образовательного предмета, по которому проводится ЕГЭ
<input type="checkbox"/> Вскрытие доставочного пакета осуществлялось НЕ в присутствии участника ЕГЭ	<input type="checkbox"/> Наличие нарушений дисциплины в аудитории

## FineReader Forms

Единый государственный экзамен - 2005  
**Бланк ответов № 1**

Заполнить геленой или капиллярной ручкой ЧЕРНЫМИ чернилами ЗАГЛАВНЫМИ ПЕЧАТНЫМИ БУКВАМИ по следующим образцам:  
А В В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Э Ю Я 1 2 3 4 5 6 7 8 9 0  
А В В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Э Ю Я 1 2 3 4 5 6 7 8 9 0

**ВНИМАНИЕ!** Данный бланк использовать только совместно с двумя другими бланками из данного пакета

Информация об участнике ЕГЭ

Фамилия \_\_\_\_\_ Имя \_\_\_\_\_ Отчество \_\_\_\_\_

Документ \_\_\_\_\_ Сервис \_\_\_\_\_ Номер \_\_\_\_\_ Пол:  Ж.  М.

Резерв - 1 \_\_\_\_\_ Резерв - 2 \_\_\_\_\_ Резерв - 3 \_\_\_\_\_ Факт выхода из аудитории во время экзамена

**Замечания участника ЕГЭ по процедуре проведения ЕГЭ**

Заполнение НЕ ОБЯЗАТЕЛЬНО.

Отметьте  замечания по проведению экзамена:

<input type="checkbox"/> Организация доставки участника в пункт проведения ЕГЭ при самостоятельном времени в пути более 1 часа	<input type="checkbox"/> Присутствие в аудитории преподавателя образовательного предмета, по которому проводится ЕГЭ
<input type="checkbox"/> Вскрытие доставочного пакета осуществлялось НЕ в присутствии участника ЕГЭ	<input type="checkbox"/> Наличие нарушений дисциплины в аудитории

- **Бланком** называется стандартный лист бумаги, на котором размещается постоянная информация и отведено место для переменной.
- Сложность состоит в том, что необходимо распознать написанные от руки символы, довольно сильно различающиеся у разных людей.
- Кроме того система должна определить, к какому полю относится распознаваемый текст.



## Системы оптического распознавания форм

- Для обработки бланков предназначено специальное приложение **FineReader Forms**.
- Для распознавания содержимого бланка необходимо предварительно создать шаблон формы.

### Сервис/ Шаблоны

- Шаблон используют на этапе сегментации. Сегментация в данном случае состоит в наложении шаблона.
- Положение шаблона корректируется в соответствии с тем, насколько ровно был размещён бланк при сканировании.
- Заключительный этап состоит в распознавании содержимого бланка.

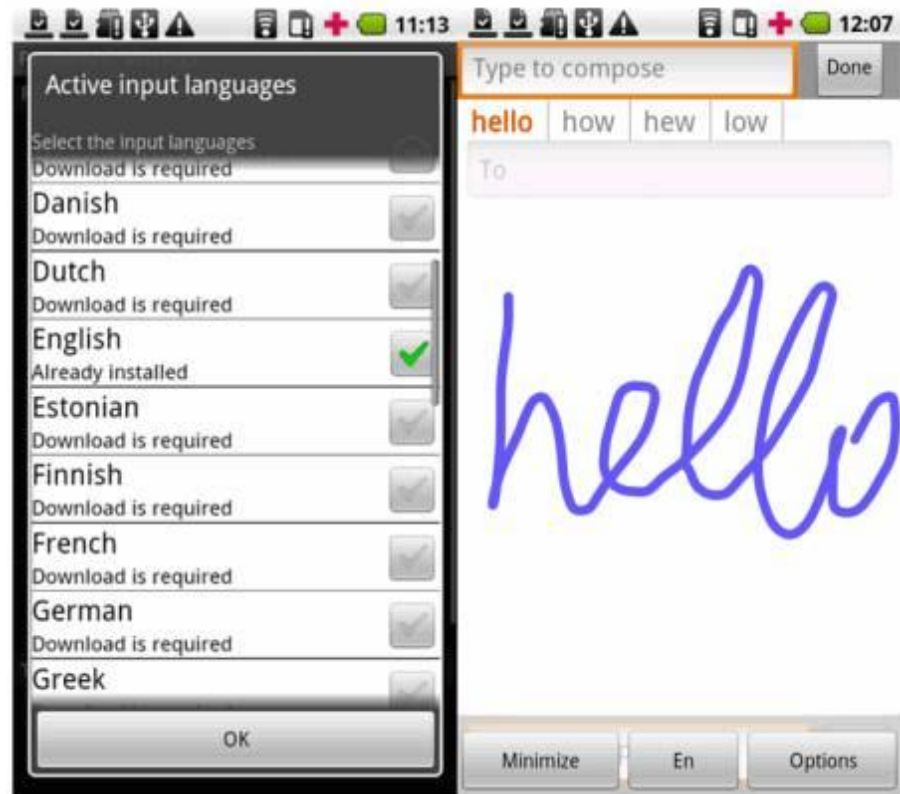
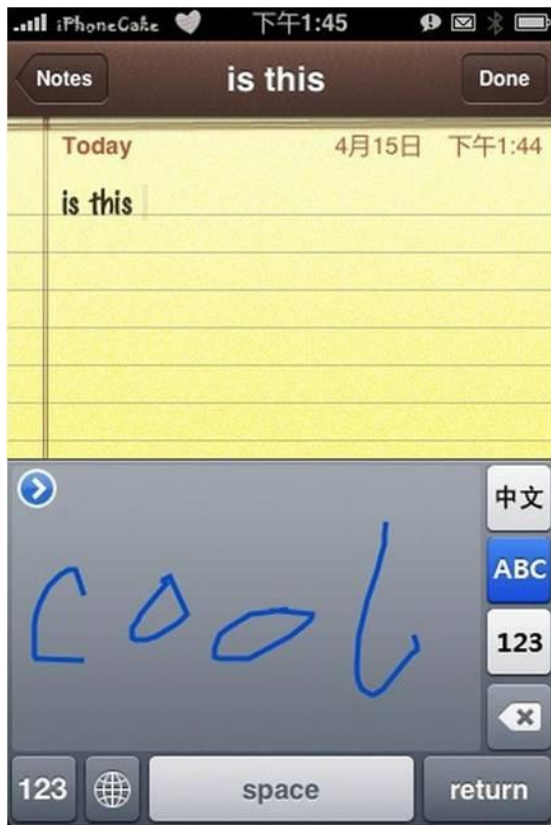
## Системы распознавания рукописного текста


С появлением первого карманного компьютера Newton фирмы Apple в 1990 году начали создаваться системы распознавания рукописного текста. Такие системы преобразуют текст, написанный на экране карманного компьютера специальной ручкой, в текстовый компьютерный документ.





# Системы распознавания рукописного текста



The background features a central point from which several colorful rays (green, orange, blue, grey) emanate outwards. Faint binary code (0s and 1s) is scattered across the background, and a small line graph with two lines (one orange, one blue) is visible on the right side.

# **Программы оптического распознавания текста**





# Программы оптического распознавания документов



Для ввода сканеров можно использовать с буферной флешкой или карту памяти мобильного телефона программы распознавания сим

Одной из наиболее популярных программ такого типа является ABBYY FineReader





# Принцип работы сканера

Принцип работы сканера состоит в следующем: в результате преобразования света получается электрический сигнал, содержащий информацию об активности цвета в исходной точке сканируемого изображения. После оцифровки аналогового сигнала в АЦП цифровой сигнал через аппаратный интерфейс сканера идет в компьютер, где его получает и анализирует программа для работы со сканером. После окончания одного такого цикла (освещение оригинала — получение сигнала — преобразование сигнала — получение его программой) источник света и приемник светового отражения перемещается относительно оригинала.

## ПРИНЦИП РАБОТЫ СКАНЕРА

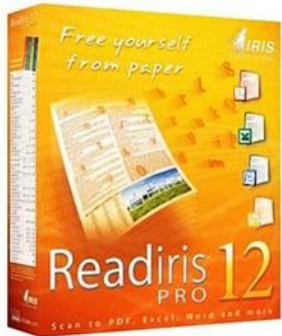
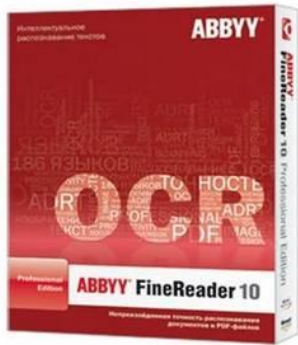


# Программы распознавания текста



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

Современная OCR должна уметь многое: распознавать тексты, набранные не только определенными шрифтами, но и самыми экзотическими, вплоть до рукописных. Уметь корректно работать с текстами, содержащими слова на нескольких языках, корректно распознавать таблицы. И самое главное — корректно распознавать не только четко набранные тексты, но и такие, качество которых, мягко говоря, далеко от идеала. Например, текст с пожелтевшей газетной вырезки или третьей машинописной копии. Само собой, распознать текст — это еще полдела. Не менее важно обеспечить возможность сохранения результата в файле популярного текстового (или табличного) формата — скажем, формата Microsoft Word.



Free Online OCR



# OCR CUNEIFORM

- Это **бесплатная** программа сканирования и распознавания текста российского разработчика Cognitive Technologies.
- **OCR CuneiForm** обеспечивает быстрое, удобное и качественное распознавание текста с сохранением исходного вида документа. Поддерживается распознавание с более 20 языков, среди них русский, украинский, английский, немецкий, французский, испанский, итальянский, португальский, шведский, финский, сербский, хорватский, польский, а также распознавание смешанного русско-английского текста.



# ➔ ABBYY FineReader

- Популярная проприетарная программа распознавания текста компании ABBYY
- Программа производит распознавание текста с более **180 языков**, для **38** из них предусмотрена встроенная проверка орфографии. Начиная с версии **Professional**, распознаются иврит, японский, тайский, китайский языки. Finereader открывает файлы графических форматов (TIFF, JPG, PFD, PNG и др.) в том числе **DjVu** – компактный формат для хранения отсканированных документов, книг.



# Окно программы FineReader

The screenshot shows the ABBYY FineReader 5.0 Office Try&Buy interface. The window title is "Default - ABBYY FineReader 5.0 Office Try&Buy - [1 - Text]". The menu bar includes "Файл", "Правка", "Вид", "Пакет", "Изображение", "Процесс", "Сервис", "Окна", and "Справка". The toolbar contains icons for file operations and a language dropdown set to "Русско-Английский". The font settings are "Arial", size "26". The main toolbar has four buttons: "Scan&Read", "Сканировать", "Распознать", "Проверить", and "Сохранить".

Annotations on the left side of the image:

- Строка меню** (Menu bar) points to the top menu bar.
- Панели инструментов** (Toolbars) points to the toolbar area containing the "Scan&Read" and "Сканировать" buttons.
- Текущий пакет страниц** (Current page pack) points to the page pack view on the left showing a thumbnail of the scanned page.

The main workspace is divided into three panels:

- Left panel:** A vertical toolbar with various icons for text manipulation.
- Middle panel:** A preview of the scanned page with a "Блочная структура текста" (Text block structure) overlaid. The structure consists of numbered green boxes (1-17) identifying different text blocks on the page.
- Right panel:** A text view showing the "Результат распознавания" (Recognition result). The text is displayed in a light blue font, indicating the recognized content.

At the bottom of the interface, there are navigation buttons and a language dropdown set to "Русский".

# ➔ Процесс обработки FineReader

- **Сканирование** (сканер, цифровой фотоаппарат, цифровая видеокамера).
- **Сегментация** - выделение блоков на изображении.
- **Распознавание** – неоднозначно опознанные символы выделяются цветом.
- **Проверка ошибок**- можно провести проверку грамматики.
- **Сохранение** результатов в виде отформатированного или неотформатированного документа, или прямой передачи в другое приложение - WORD, Excel в буфер обмена Windows.



# ➔ OmniPage

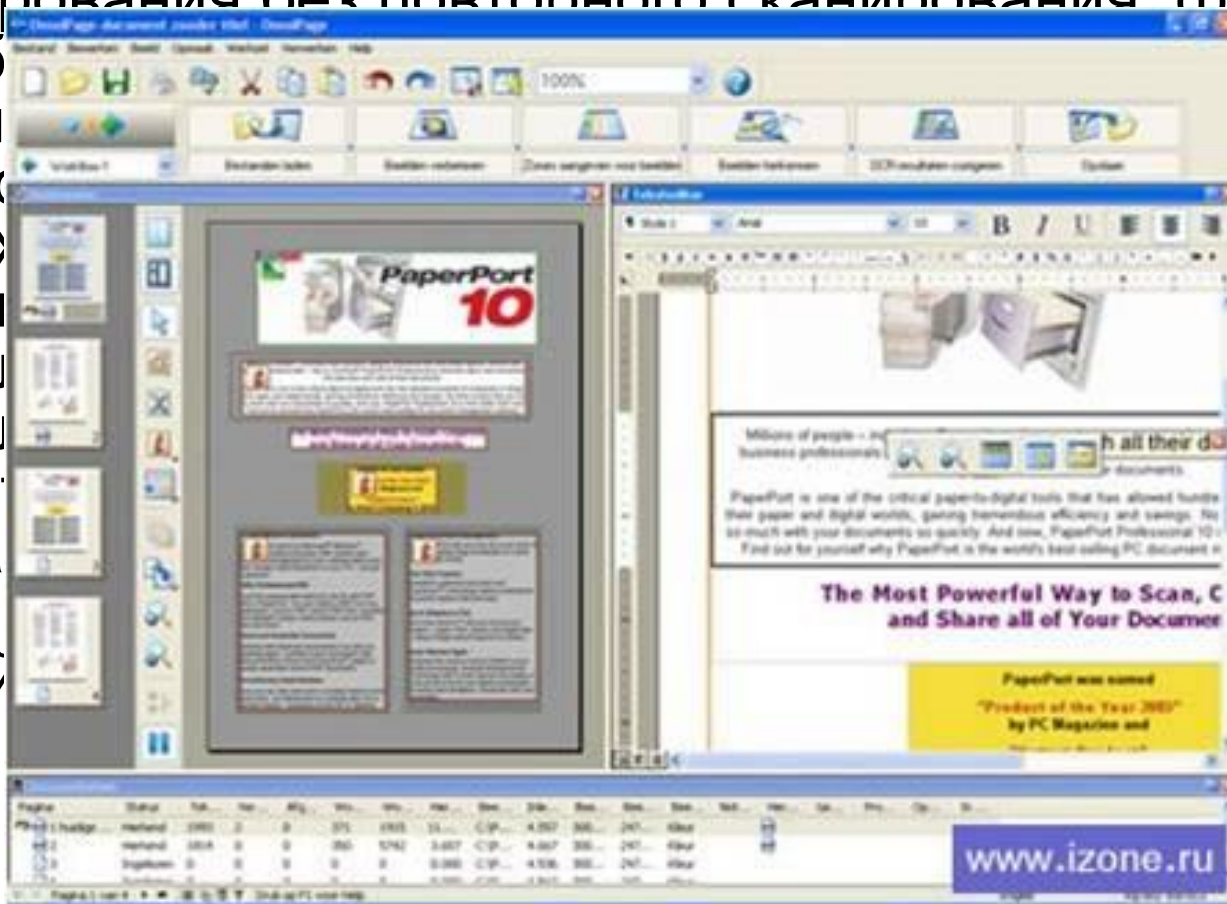
- Популярная программа распознавания текста **российской компании АВВУУ**
- Программа отличается высокой скоростью и точностью распознавания. Распознаются более **120** языков с различными алфавитами: **латинский, греческий алфавиты, кириллица, китайский, японский и корейский** языки. Как и FineReader, OmniPage уверенно распознает документы, полученные с помощью цифровых камер с помощью технологии коррекции изображения "3D Correction".



# ➔ OmniPage

- В программе присутствуют удобные инструменты обработки изображений, повышенное качество сканирования без повторного сканирования: функция

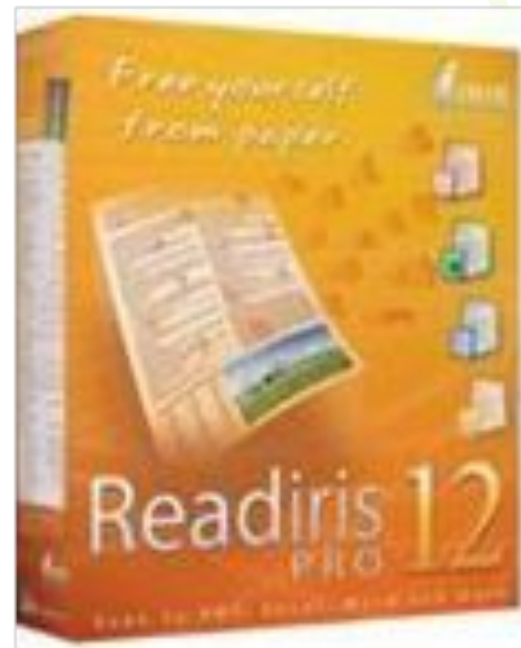
преоб  
докум  
Desktop  
других  
компл  
неско  
позво  
редак  
вариа  
обрат  
любой



е  
Google  
айла (и  
м. В  
ся  
Converter -  
DF в  
ощенный  
ет  
чески  
мат PDF.

# ➔ Readiris

- Программа сканирования и распознавания текста **компании I.R.I.S.**
- Поддерживается распознавание текста с более **120 языков** распознавания, включая русский, а также ближневосточные языки - **арабский, иврит, фарси** (в версии Middle-East) и **японский, китайский, корейский** (в версии Asian). Есть версия Readiris для **Macintosh**.
- Вместе с поддержкой распознавания популярных форматов картинок, распознаются файлы **PDF** и **DjVu**.



# Readiris



Содержит региональные пакеты для распознавания азиатских языков и языков среднего востока.

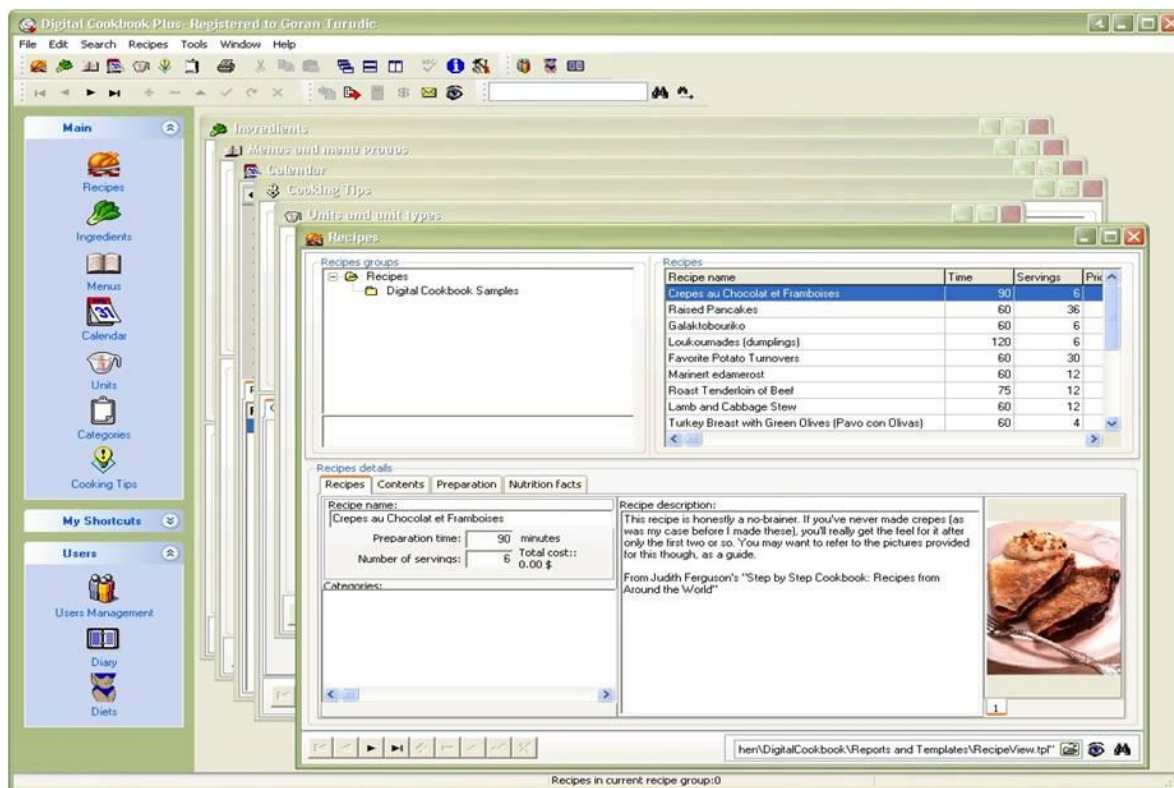


# Kirtas Technologies Arabic OCR

Может распознавать арабские и английские символы на одной странице.

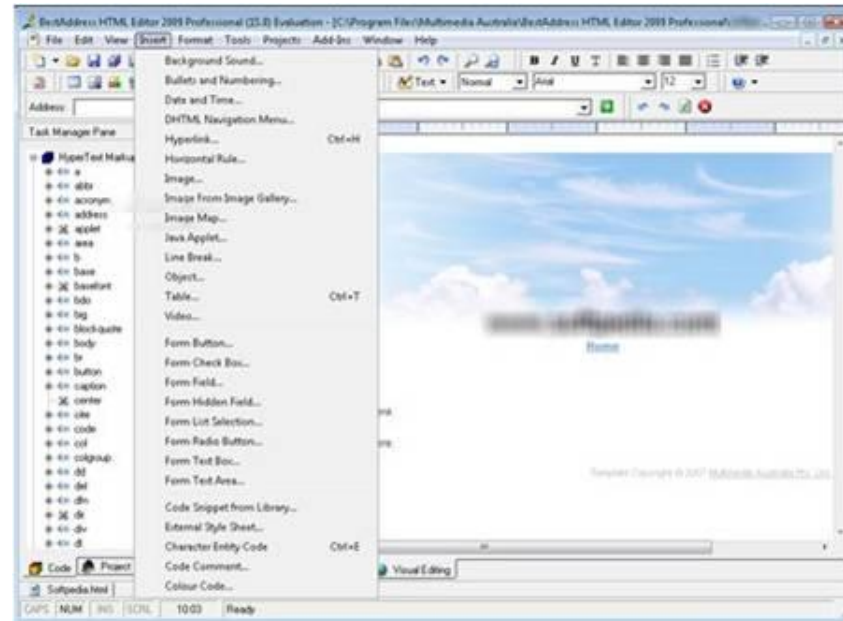


# ➔ Zonal OCR



Помогает автоматизировать извлечение данных из компьютерных изображений.

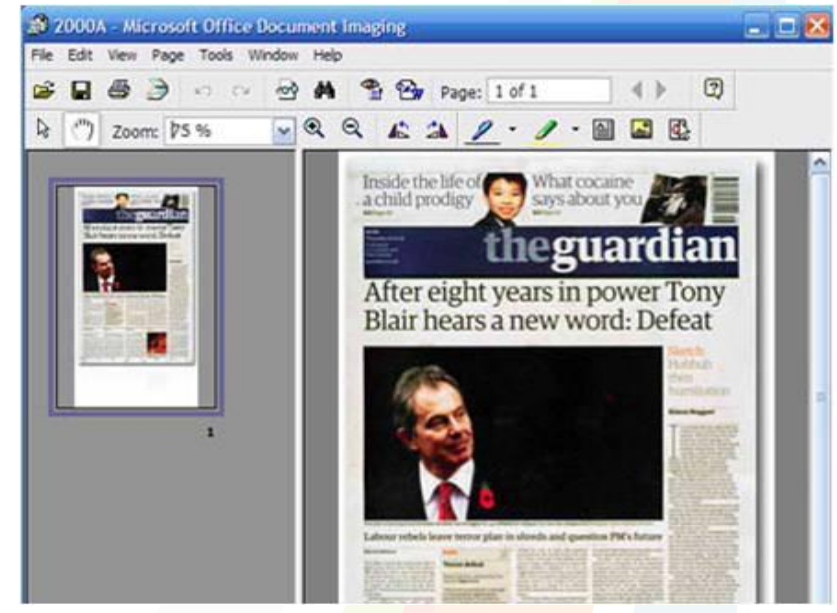
# Brainware



Извлечение данных из документов и их обработка — например, счета, извещения, накладные и платёжки

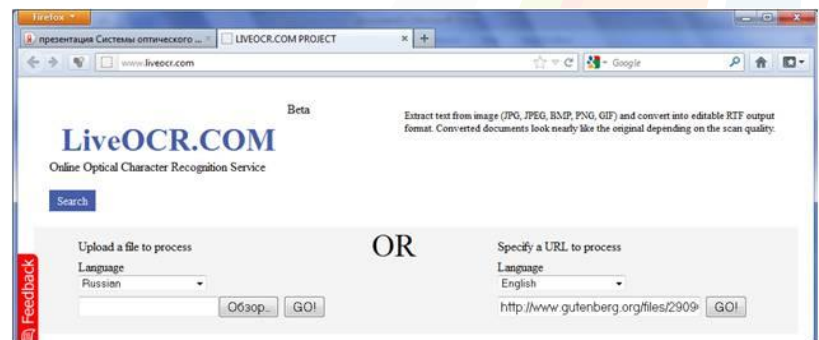
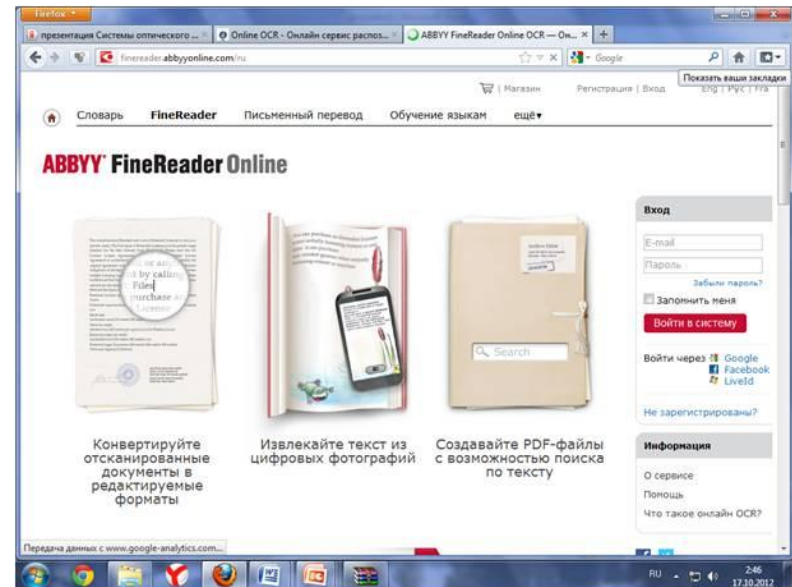
# ➔ Microsoft Office Document Imaging

- Программа распознавания текста компании **Microsoft**
- Программа Document Imaging способна работать только с **двумя** языками: английским и языком локализации самого MS Office. Для поддержки других языков необходимо дополнительно устанавливать пакет **Multilingual User Interface (MUI)**. **OCR** настроек в программе практически нет, программа в автоматическом режиме поддерживает распознавание типа и размера шрифтов, картинок и простых таблиц.





➔ Существует также системы On-line распознавания текста:  
**Online OCR** и **ABBYY FineReader Online**  
(<http://www.onlineocr.ru> , <http://finereader.abbyyonline.com>,  
<http://www.liveocr.com/> )



## Подведение итогов урока

1. В чем состоят различия в технологии распознавания текста при использовании растрового и векторного методов?
2. Для чего предназначены программы оптического распознавания документов?



 **Домашнее задание:**

- П. 2.8 стр. 71-73