Сканеры и программное обеспечение распознавания символов

Анна Виноградова

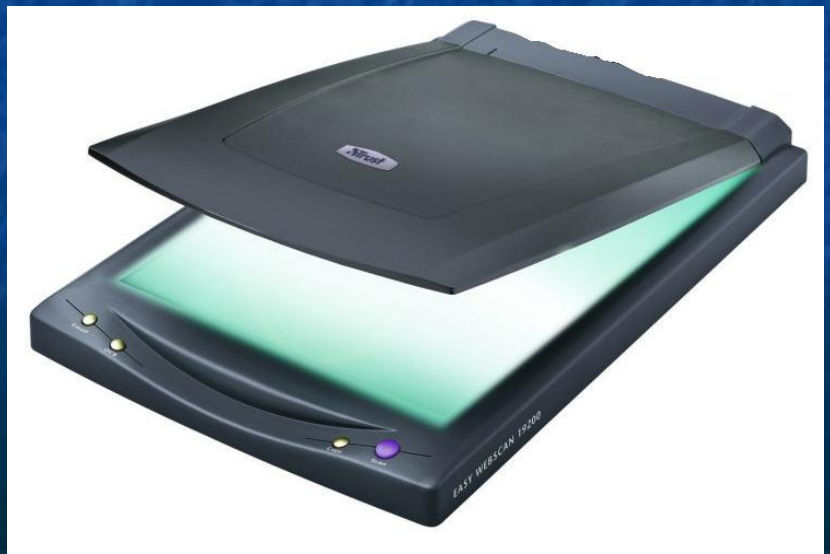
- Сканер - оптико-электронное устройство для ввода в компьютер графических изображений.
- Сканер создает оцифрованное изображение документа и помещает его в память компьютера.

Виды сканеров

- Планшетные
- Протяжные или роликовые
- Планетарные или книжные

Планшетные

- Планшетный сканер - сканер, в котором оригинал кладется на стекло и сканируется при помощи подвижной линейной ПЗС матрицы.
- Предназначены для ежедневного сканирования небольшого объёма фотографий, слайдов или документов.
- простота конструкции, ручная обработка документов, низкая производительность.



Протяжной или роликовый сканер.

- двустороннее сканирование (дуплекс)
- подсветка оригинала разными цветами для отсечки цветного фона
- система компенсации неоднородного фона
- модули динамической обработки разнотипных оригиналов
- надпечатывание отметки о том, что документ прошел обработку

Планетарный или книжный сканер.

- предназначен для сканирования скреплённых документов, периодических изданий и книг.
- бесконтактный метод сканирования
- большая производительность при оцифровке книг и сшитых оригиналов.

Основные характеристики сканеров

- Разрешение (Resolution) - число точек или растровых ячеек, из которых формируется изображение, на единицу длины или площади.
- Измеряется в "точках на квадратный дюйм" (DPI, dots per inch).
- Типовое разрешение промышленных сканеров - 200-300 DPI.

- Разрядность цвета - количество разрядов каждого пикселя в цифровом изображении. Описывает максимальное количество цветов, воспроизводимое сканером в виде степени числа 2.
- Время сканирования измеряется в страницах в минуту.
- Формат сканируемого документа - A3/A4.
- Интерфейс передачи данных могут быть различны (к COM или USB порту, к SCSI карте и др.).

Программное обеспечение сканирования

Существует три категории ПО сканирования:

- ПО сканирование малых объёмов документов
- ПО сканирования больших объёмов документов
- ПО для специальных задач сканирования

ПО сканирование малых объёмов документов.

- Применяется при домашнем и офисном сканировании.
- Используется практически любое программное обеспечение, совместимое со стандартом TWAIN и поддерживающее функцию сканирования.
- TWAIN - является стандартом для прикладного программного интерфейса (API) таких периферийных устройств, как сканеры.

Примеры ПО сканирования малых объемов:

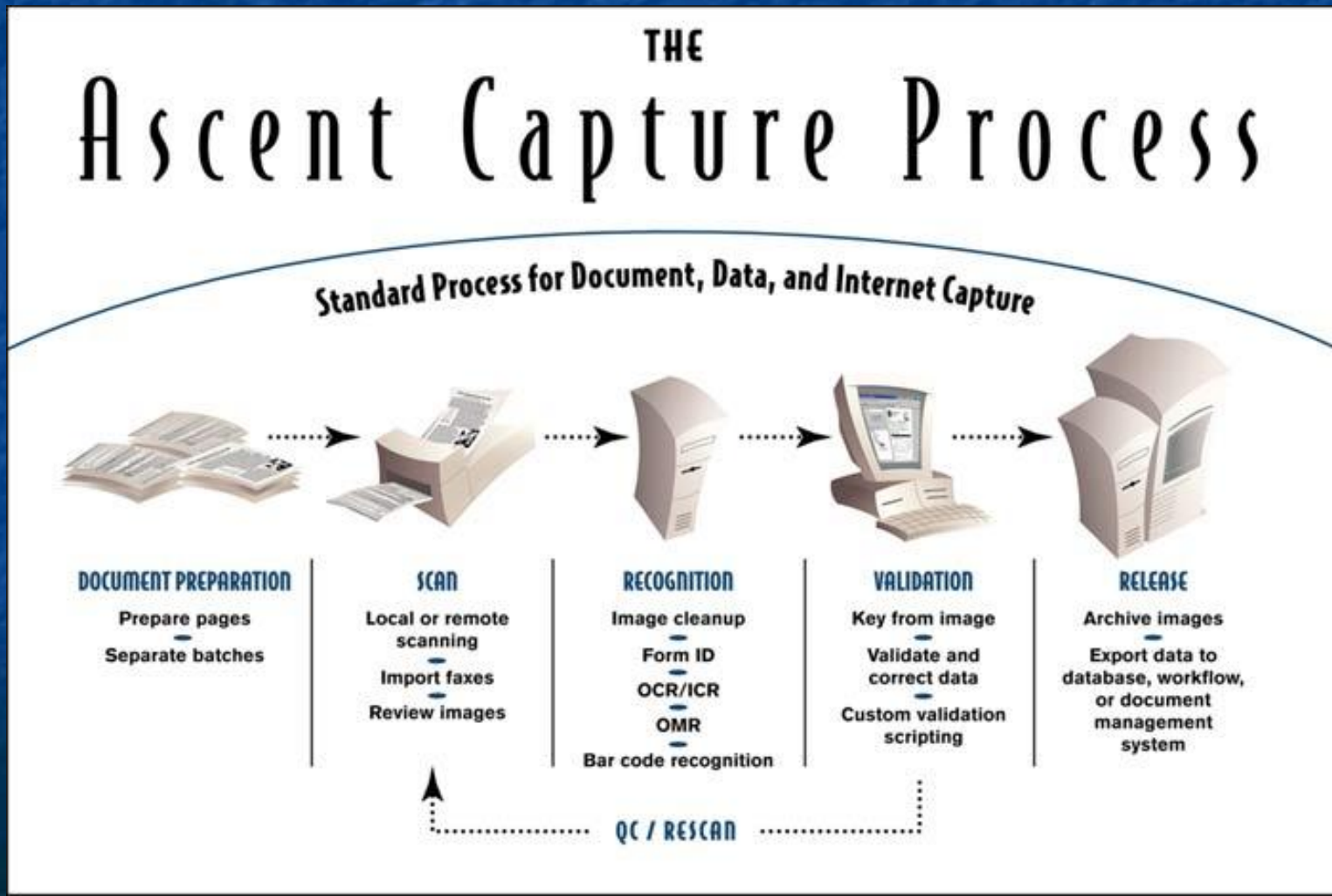
- ABBYY FineReader
- Adobe PhotoShop
- Cognitive Cuineiform
- Microsoft Photo Editor
- ACDSEE

ПО сканирования больших объёмов документов.

- Применяется при промышленном сканировании.
- При поточном сканировании один сканер ежедневно может обрабатывать до 50.000 и более документов.
- Для программного управления сканерами используется промышленный стандарт ISIS (ISIS - Image and Scanner Interface Specification).

Примеры применяемого ПО:

- Kofax Ascent Capture
- Captiva InputAccel



ПО для специальных задач сканирования.

- Применяется при планетарном, высококачественном сканировании.
- Разрабатывается с учётом специфики сканирующего устройства

Применяемое ПО:

- ПланСкан BSC-2
- RZ ProScan Book (Minolta PS7000 edition)
- Zeutschel OmniScan

Что такое системы распознавания?

- Необходимо выполнить сканирование бумажных документов и распознать их содержимое с помощью специальных программ (Optical Character Recognition - OCR).
- Системы оптического распознавания символов предназначены для автоматического ввода печатных документов в компьютер. Обработка изображения OCR-системой включает в себя анализ графического изображения и распознавание каждого символа.

Процессы анализа макета страницы:

- определение областей распознавания
- определение таблиц
- определение картинок
- выделение в тексте строк и отдельных символов

Точность распознавания

- OCR-системы могут достигать наилучшей точности распознавания — свыше 99,9% для чистых изображений.
- Если имеется приблизительно 1500 символов на странице, то даже при коэффициенте успешного распознавания 99,9 % получается одна или две ошибки на страницу. В таких случаях на помощь приходит метод проверки по словарю.
- Но это все равно не позволяет исправлять 100 % ошибок, что требует человеческого контроля результатов.

Причины ошибок при распознавании

- Грязные изображения
- Неаккуратное сканирование, связанное с «человеческим фактором»
- Если документ был ксерокопирован, нередко возникают разрывы и слияния символов.
- Страница, расположенная с нарушением границ или перекосом, создает немного искаженные символьные изображения, которые могут быть перепутаны OCR.

Спасибо за внимание!