

OCR2PDF
Production OCR

Why OCR2PDF?

- Once you have decided to scan your paper documents, why not go all the way and turn paper into text, instead of just images?
- We often have to browse in sequence tens of pages, if not even hundreds or thousands, spending value time just to find “that” useful information
- What if the same search would take now only a few seconds, or less?
- What if you could also perform “transversal” researches that might provide additional information that would have been undiscovered, laying hidden in the paper documents or in the scanned images?

Product Concept

- OCR2PDF is a professional solution that converts high volumes of non-searchable text images into full-text searchable PDF files
- For any organizations that manage critical paper documents, OCR2PDF allows making strategic business information stored on paper easily accessible and reusable to anybody who needs it
- Whether needing to integrate volumes of paper documentation and digital data, OCR2PDF definitely bridges the gap between paper and digital workflows
- OCR2PDF is a production tool, a “black box” able to process
 - High volumes of image documents
 - With high recognition speed
 - With high recognition accuracy

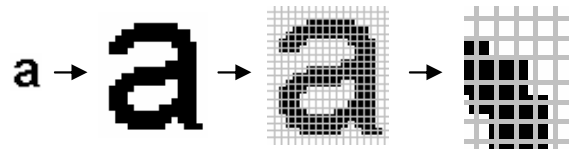
Functional Architecture

- Input:
 - OCR2PDF can process TIFF images or non-searchable PDF
- Process:
 - OCR2PDF recognizes the text and creates a PDF file that contains the page image and the recognized text
 - The page image is contained above the text in a two-layered PDF file
 - Process can be activated on-fly or can be run in batch mode through an integrated scheduler
- Output:
 - PDFs delivered by OCR2PDF can be searched with Acrobat or the free Acrobat Reader

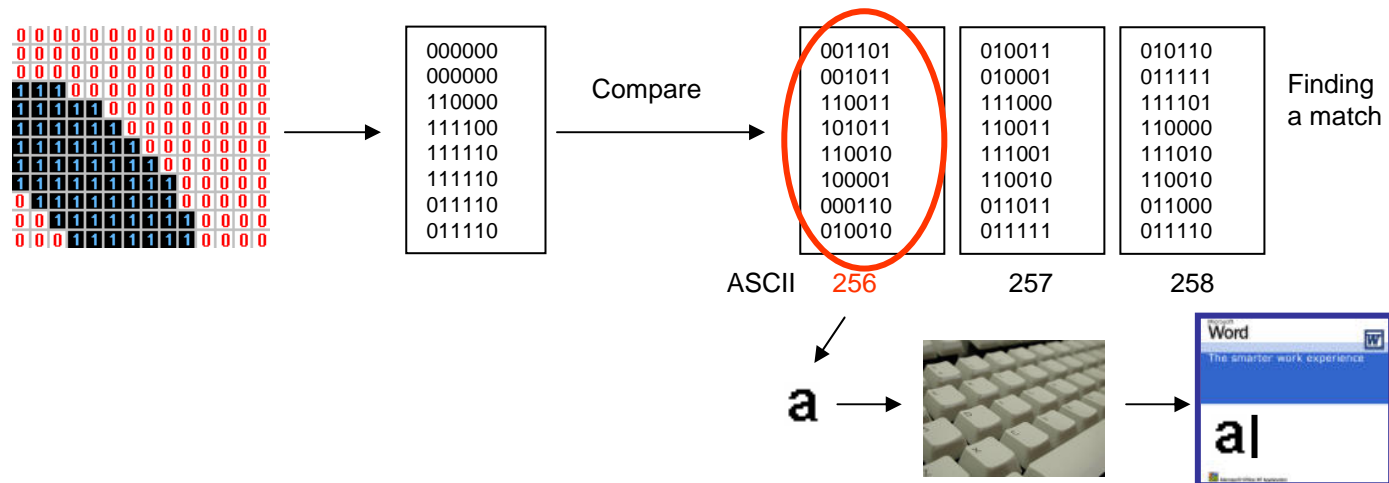


How does OCR work?

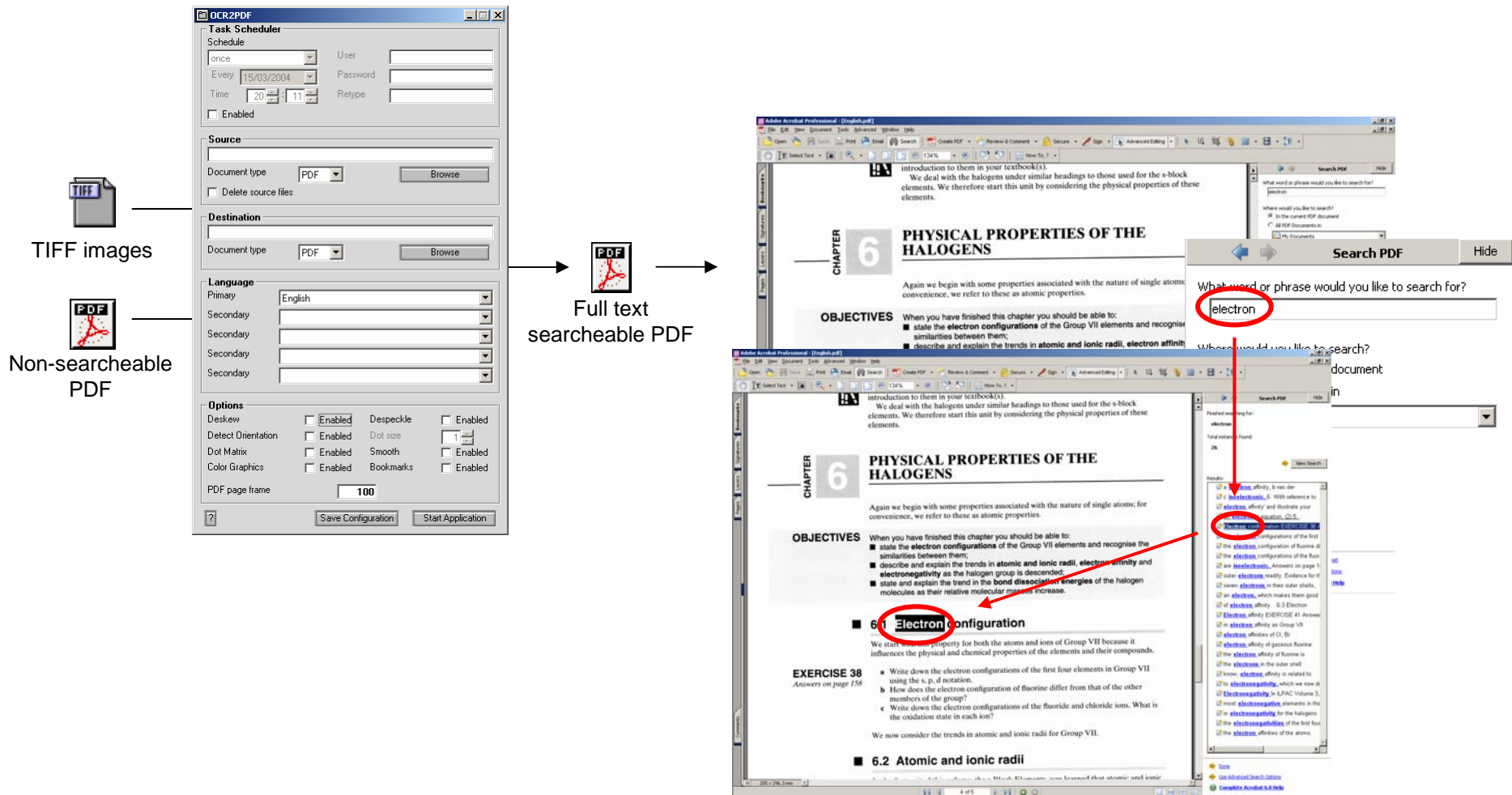
- Step 1: scanned images are a meaningless cloud of intense points, “pixels”, on a lighter background. If no bi-tonal is available, intelligent binarization routines convert color and greyscale images into black-and-white images.
- Step 2: a line segmentation is done. It consists of slicing a page of text into its different lines. This step also analyses line skew, interline spacing and drop letters, and separates touching lines. The word and character segmentation phases isolate one word from another and separate the various letters of a word:



- Step 3: the software analyzes the segmented characters as human beings unconsciously do: a number of features are extracted (strokes, loops, holes, nodes, angles, etc.), and checked against a predefined source of knowledge:



Workflow



Note: program interfaces are subject to changes

Technical Info

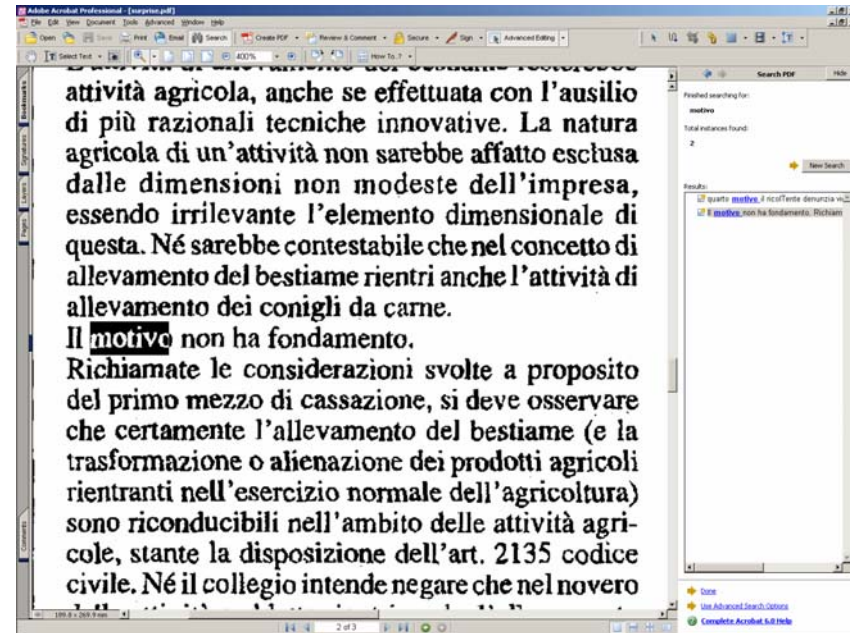
- **Speed and accuracy**
 - Four times faster than any other OCR. With an accuracy that surprises
 - Running on a 2 GHz Pentium IV PC, OCR2PDF can recognize up to 3000 characters per second.
 - Recognizing an A4 page can take no more than one to five seconds per page
 - Unless you dispose of a high-volume document scanner, the OCR phase can be faster than the scanning
 - Given correct use of capture software and OCR2PDF, you should get a recognition rate above 99% on most documents, including low-quality documents such as photocopied letters or faxes
- **Automatic deskew**
 - Can detect the document orientation and text disalignment
 - Pages are rotated and deskewed for optimal accuracy
- **Recognition of color documents**
 - Recognizes color documents and texts on color backgrounds
- **Multi-language support**
 - Can recognize up to 104 different languages (dictionaries)
 - An optional add-on is also available for Japanese, Korean, Traditional Chinese and Simplified Chinese
 - Up to 5 languages can be simultaneously active during each recognition
- **Bookmarks**
 - Associates a bookmark to any title, picture, or table. If input file already has meaningful bookmarks, they can be imported into output file
- **Optimized PDF management**
 - An exclusive algorithm has been implemented to manage PDF files with thousands of pages keeping performances at the highest rate
- **Report file**
 - A report file is created for each process, summarizing all job settings

System Requirements

- PC
 - Windows 2000, XP Professional
 - PC Pentium IV - 3 GHz or higher
 - 512 MB RAM minimum (1 GB or higher recommended)
 - SuperVGA card, min 1024x768 - 24 bit
- Available Languages
 - English
 - Versions localized in other languages can be provided on demand

Application Samples

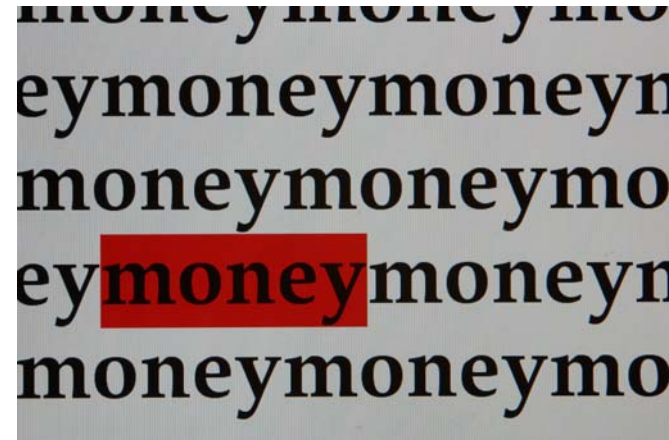
- Legal
 - Lawsuit files
- Finance
 - Contracts
- Education
 - Contents extractions from training courses
- Government organizations
 - Paperwork, laws, communications
- Manufacturing
 - Manuals, product documentation, researches, reports
- Healthcare
 - Forms
- Human Resource Departments - Recruiting Agencies
 - C.V.'s
- Service Bureaus
 - to offer new professional services on Customer's documents



- After processing the files with OCR2PDF, searchable PDFs can be indexed and archived through a Document Management System (e.g. → Ad Hoc Digicarta)

Business Benefits

- Time savings
 - OCR2PDF allows recognition of huge document volumes with speed and accuracy, with an efficient workflow organization, and with the optimal use of all involved resources (PC, scanner, etc.)
 - Running the OCR process overnight lets the user focus during the working day on core activities: capture, search, reading contents, organizing, etc. That means incredible time saving, to guarantee the immediate return of your investment
- Productivity
 - OCR2PDF delivers unmatched combination of speed and accuracy
 - It is the ideal solution for professional users who have to process high document volumes, and need to achieve the maximum recognition rate
- Versatility of PDF
 - OCR2PDF generates PDF files fully searchable, containing both images and text. In the full-text PDF you can search a whole word or part of a word
 - If you have Adobe Acrobat, thanks to the “Catalog” capability it is possible to create full-text indexes of entire PDF collections
 - The output PDF are ideal for archiving and for using in document management systems
- Reuse contents of paper documents
 - If document contents need to be re-used, OCR2PDF can generate RTF files as well, that can be edited in any wordprocessor: texts, photos, drawings, tables are correctly positioned, and any further modification and fine tuning of the recognized document is easy and quick
 - Thanks to OCR2PDF you do not have anymore to retype from scratch full documents received on paper





Ad Hoc Consulting S.r.l.
Via della Moscova 46/3
20121 Milano
Italy
Tel.: +39-02-97069301
Fax: +39-02-87399063

adhoc@adhoc-online.com
www.adhoc-online.com

Ad Hoc
CONSULTING