

# OCR2PDF



The Professional OCR

Although the first research and development on Optical Character Recognition (OCR) began more than 30 years ago in “Artificial Intelligence” labs and as popular as scanners of all types may have become, this technology is still unknown by some people who could use it for their document entry applications. The first systems were pretty slow, not all that accurate and in most cases limited to the recognition of special fonts (OCR-A, OCR-B) used to encode data on cheques. But in the field of OCR, as in other fields of computer science, huge progress was and is being made, and the technology has become accurate, fast and stable. Ad Hoc offers today a professional OCR product that is best-in-class in the market, in performances as well in accuracy. If you really want to benefit of the leading recognition technology, you have met the OCR that works.

Once you have decided to de-materialize your 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. The OCR2PDF purpose consists in processing high volumes of text images and generate effortless collections of text documents, fully searchable. The same search takes now only a few seconds, or less! Furthermore, you can now also perform “transversal” researches that may output additional information that would have been undiscovered, laying hidden in the paper documents or in the scanned images.

### BEST-IN-CLASS RECOGNITION TECHNOLOGY

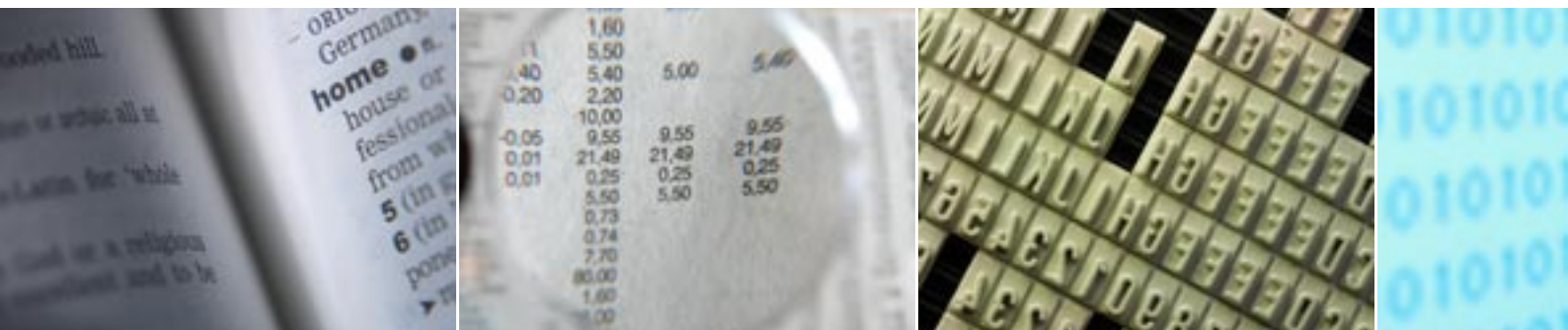
OCR2PDF is a highly professional solution that converts text images into full-text searchable PDF files.

The OCR2PDF technology is based upon a font-independent feature extraction, complemented by self-learning techniques derived from a proprietary neural network. “Font-independent feature extraction” means a topological analysis is performed, with omnifont recognition (you can recognize virtually all fonts). “Self-learning techniques” refers to the autolearning which is the result of linguistic context analysis: the system learns font shapes it did not know beforehand because the linguistic context points it in the right direction. The “neural network” is the decision model used to organize it all.

When the document is scanned, the document image is only 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. The OCR software extracts text information from the black-and-white pixels, recognizes the shapes and assigns characters.

Firstly, 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.

The actual character recognition extracts characteristics out of each isolated shape and assigns a symbol. 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.



### NEEDS ALWAYS ACTUAL

The need for document capture is actually much bigger than you would think at first.

An American study by IDC indicated that 55% of all typing done on PCs is actually retyping of data already available on paper.

Other studies indicate that, despite the many imaging applications, 90% of all information is still held on paper, a shocking fact for our computer age!

### ACCURACY

Scanning factors have an enormous influence on the image quality: if you degrade a document by scanning it badly, the recognition is bound to suffer. Secondly, make sure that the right OCR settings are enabled (e.g. do not try to read a French document with the language set to English).

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.

### SPEED

Running on a 2 GHz Pentium IV PC, OCR2PDF can recognize up to 3000 characters per second.

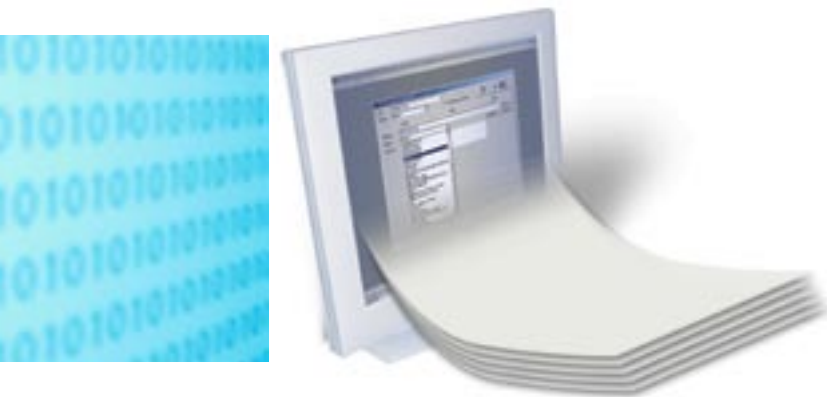
Recognizing an A4 page can take no more than one to five seconds per page. That's the kind of mileage you get out of a single workstation.

Compare the rapidity of the text conversion for a moment with the rated speed of your scanner: unless you dispose of a high-volume document scanner, the OCR phase will be faster than the scanning!

## THE SIMPLEST WORKFLOW

OCR2PDF offers several parameters to give optimal results. The typical workflow includes following steps:

1. Documents are scanned. Paper originals are converted to TIFF (single page or multipage TIFF), or to a PDF containing only page images. (It is here assumed that these operations are performed by a capture software, they are not part of OCR2PDF).  
By the way, input files do not have to be necessarily generated by a scan session: they could have been generated by any legacy application, or a fax server, or any third party sources.
2. For security reasons, user who wants to run OCR2PDF may log by inserting own username and password.
3. Time settings: OCR2PDF can be run “on fly”, or it can be scheduled to run at a precise, wished time. You can also set a frequency for processing (daily, weekly, monthly), and even what day of the week the process must start: e.g. every Wednesday at 7.30pm OCR2PDF will start automatically.
4. User can now set the system parameters: the input folder, where the input files are stored; and the output folder, where the resulting PDF will be saved. An exclusive algorithm allows optimizing the overall recognition performances in case of huge PDF files, containing hundreds or thousands of pages.
5. You can now set the typical OCR parameters: dictionaries (up to five concurrently), page deskew, bookmarks, dot removal, etc.
6. Once started, OCR2PDF saves configuration for later reuse, and a console monitor informs the user about the processing status. At the end, output PDF will be available in the chosen output folder, ready for any further use.



7. PDF files created by OCR2PDF can be read with Acrobat Reader (freely available), that allows searching words in the document body or in the bookmarks, either on a file or in a folder, or in a PC or network path.  
The quantity of meaningful and helpful information that you can pick up from your documents thanks to OCR2PDF is unbelievable!

## BEST-IN-CLASS VERSATILITY

OCR2PDF accepts as input files PDF documents or TIFF documents. Standard output is a full-text searchable PDF, although RTF (“Rich Text Format”) files can be produced as well.



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. Size of searchable PDF is only slightly bigger than the original file, but they offer much higher value to the users, since you can search and quickly find any word or combination of words, or localize sections of the document by using bookmarks.

Page analysis (or “page decomposition”) is the process whereby a page gets “broken up” in text blocks and graphics. The various blocks as occur on a scanned page are detected by the software. Intelligent routines analyze color, greyscale and bi-tonal images on an equal basis to detect which zones contain text and which zones contain graphics. Page analysis is particularly useful when columnized texts and documents with a complex page layout are OCRed. As OCR2PDF performs the OCR autonomously, the page decomposition is entirely automatic and invisible but highly accurate.

### COMPELLING VALUE PROPOSITION

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.

### APPLICATIONS AND MARKETS

Ideally suited for paper-intensive applications and industries:

- Legal (lawsuit files)
- Finance (contracts)
- Education (contents extractions from training courses)
- Government organizations (paperwork, laws, communications)
- Manufacturing (documentation, researches, reports)
- Healthcare (forms)
- Human Resources - Recruiting Agencies (resume’s)
- Service Bureaus (to offer new professional services on Customer’s documents)

## SOLUTION 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, and thanks to the "Catalog" capability of Adobe Acrobat you can 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 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 documents received on paper!

## TECHNICAL CHARACTERISTICS

### Speed and accuracy

Four times faster than any other OCR. With an accuracy that surprises.

### Automatic deskew

Can detect the document orientation and text disalignment. Page is 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.

### 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.

## PRODUCT COMPLEMENTS

- Book Builder
- Digicarta
- DIGIT
- Docs
- Docstation
- PDflow

## SYSTEM REQUIREMENTS

- Windows 2000, XP Professional
- PC Pentium IV
- Min 512 MB RAM
- SVGA card or better
- Available USB or PARALLEL port
- CD-ROM drive (installation)
- 200 MB available disk space
- Acrobat Reader

## AVAILABLE LANGUAGES

- English

**Ad Hoc**  
CONSULTING

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