

# **File by OCR Manual**

December 19, 2007

## **eDocFile, Inc.**

2709 Willow Oaks Drive

Valrico, FL 33594

Phone 813-413-5599

Email [sales@edocfile.com](mailto:sales@edocfile.com)

[www.edocfile.com](http://www.edocfile.com)

# File by OCR

## ***Purpose:***

The purpose of File by OCR is to automatically file a document by its contents. File by OCR will allow a user to extract data contained in a document, build a file naming and storing structure on the captured values and create a CSV file for importation into a document management system.

File by OCR has the capability to monitor an unlimited number of file folders that contain different document types to be processed, making it ideal for use with a copier that has a scan to file option. The program also supports Twain Scanners and has an easy to use interface that correctly places the file in the correct folder for processing.

The program relies completely on OCR technology which is not 100 percent accurate, when setting the program up the user should take this into consideration and capture enough data so that they can be assured that if the document is not found on the first search it can be found on a subsequent search.

If possible the user should consider formatting their documents so that mission critical data is placed on the document in large characters using an OCR font. For instance a work order number or invoice number, properly formatted and captured would allow the user to go directly to the document.

Unlike typical zonal OCR programs, File by OCR uses Optical Character Recognition on the entire document and then parses the data contents, allowing the user to easily capture data from multi-page documents and documents of various lengths such as sales receipts.

## System Requirements:

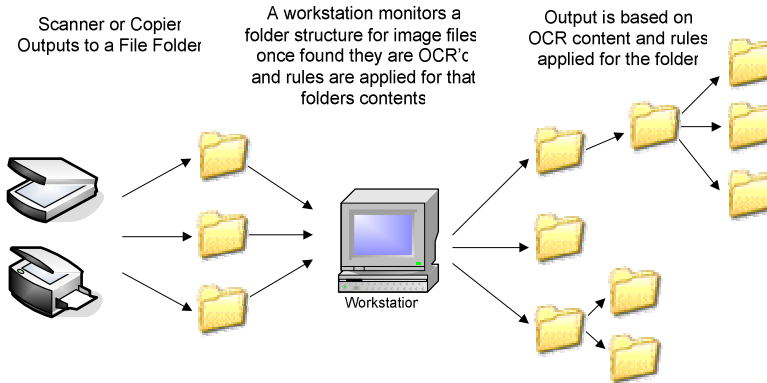
XP

Windows 2000

Microsoft Office Document Imaging

If not using a Copier that scans to a file folder on a network a Twain Compatible Scanner

## ***How it works:***



The user scans a file into a folder on the network, this folder is monitored and the file is OCR'd. Once this is completed, the data is parsed and processed based on a set of rules that relate to the input folder. For instance sales receipts could go into one folder and work orders into another. When the folder is processed the rules would be applied for receipts in all documents in folder "A" and if the document was in folder "B" the work order rules would be applied.

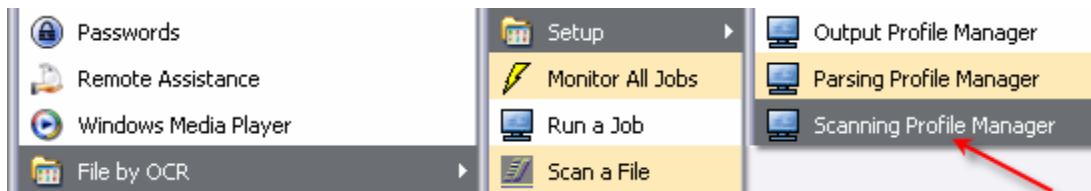
### ***Setting up a simple process:***

#### **Step 1 – Acquire an image to process**

If using a Copier create an Output folder on the network and have the image files arrive as tiff files, and go to Step 2.

#### **Setting up Scanning**

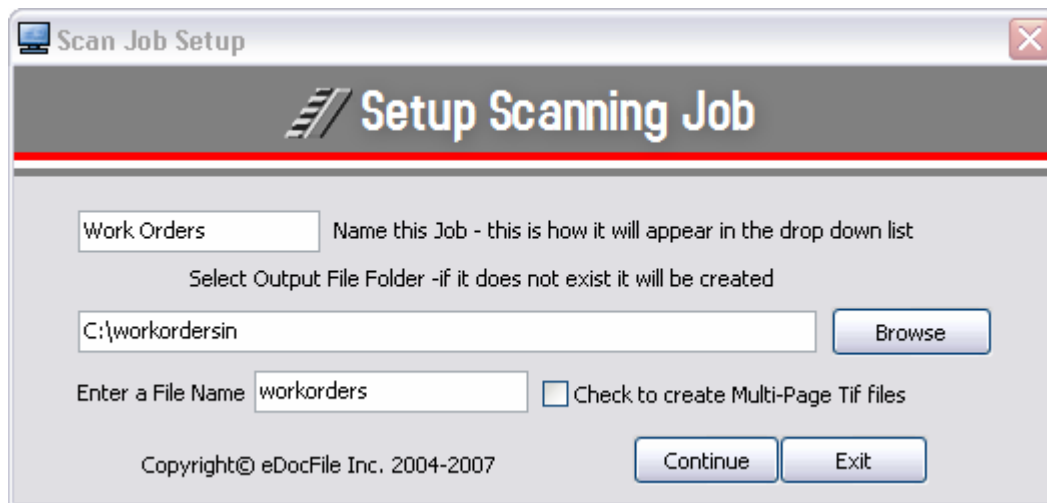
Click on Start/All Programs/File by OCR/Setup/Scanning Profile Manager



A Window will open



Click on New



Enter a name for the type of document to be processed. This name will appear in a drop down list when scanning. An unlimited number of names can be used, so it can be used to create an index value, for instance if you want to index the work order by the person that did the work, you could have it Bobs Work Orders and scan all of his work orders with this job. Or if processing receipts from different departments it could be called Pro Shop Dining Room or Mens Grill if they came from a club.

In most cases the documents to be processed will be single pages, if multi-page files are to be created place a check in the “Check to create Multit-Page Tif files” box.

Click on Continue to finish setting up scanning.

### **Scanning a File**

Click on Start/All Programs/File by OCR/Scan a File

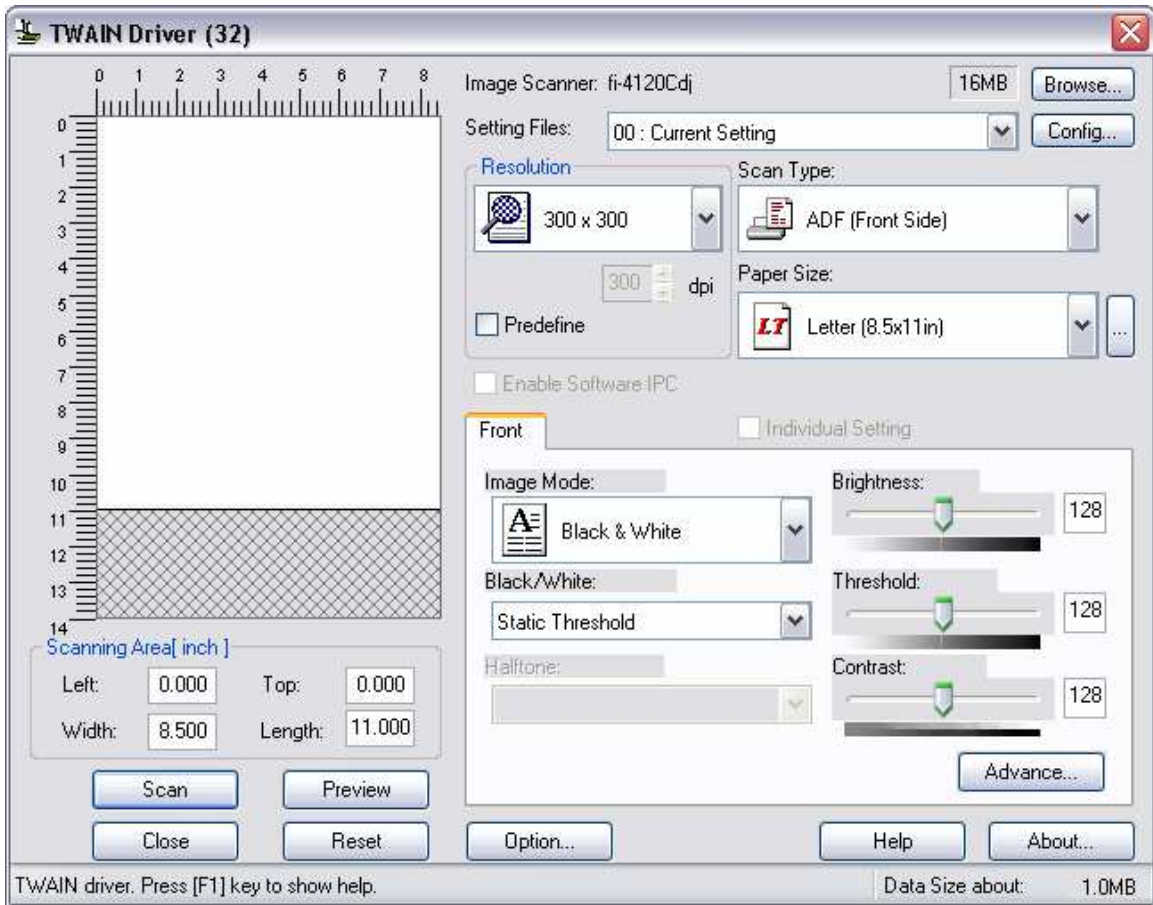


A new Window will open



Select the Document Type from the Drop Down List and click on Scan File

Note: The scanner interface will be the one for the default twain device on the PC. The menu shown here is for a fujitsu 4120c.



Check the paper size and set the resolution to 300 X 300 and the image mode to Black and White, and click on Scan. When finished scanning click on Close. When scanning a multi-page document, place an entire document in the scanner, one at a time. When it is finished scanning the document, click on scan again. When scanning single page document it makes no difference if they are scanned in a batch or not, they will all be separated into single files.

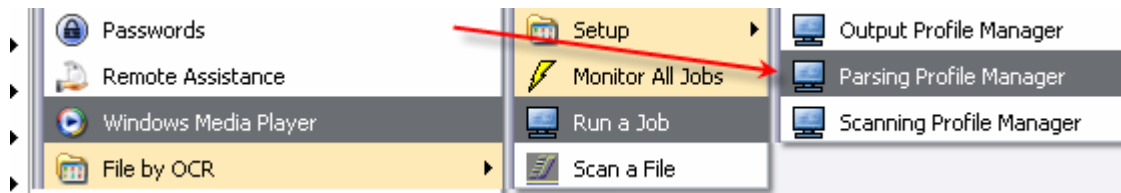


When finished scanning click on Exit

## Step 2 – Setting up parsing

This is the most difficult part of the setup process, as the OCR text may not always produce consistent results. To compensate for this the parsing is very flexible and if it cannot parse the documents correctly, contact eDocfile and supply us with some sample images. There is no one best way to parse the data, some items can be parsed many different ways.

To begin click on Start/All Programs/File by OCR/Setup/Parsing Profile Manager.



A Window will open



Click on New And a New Window Will Open

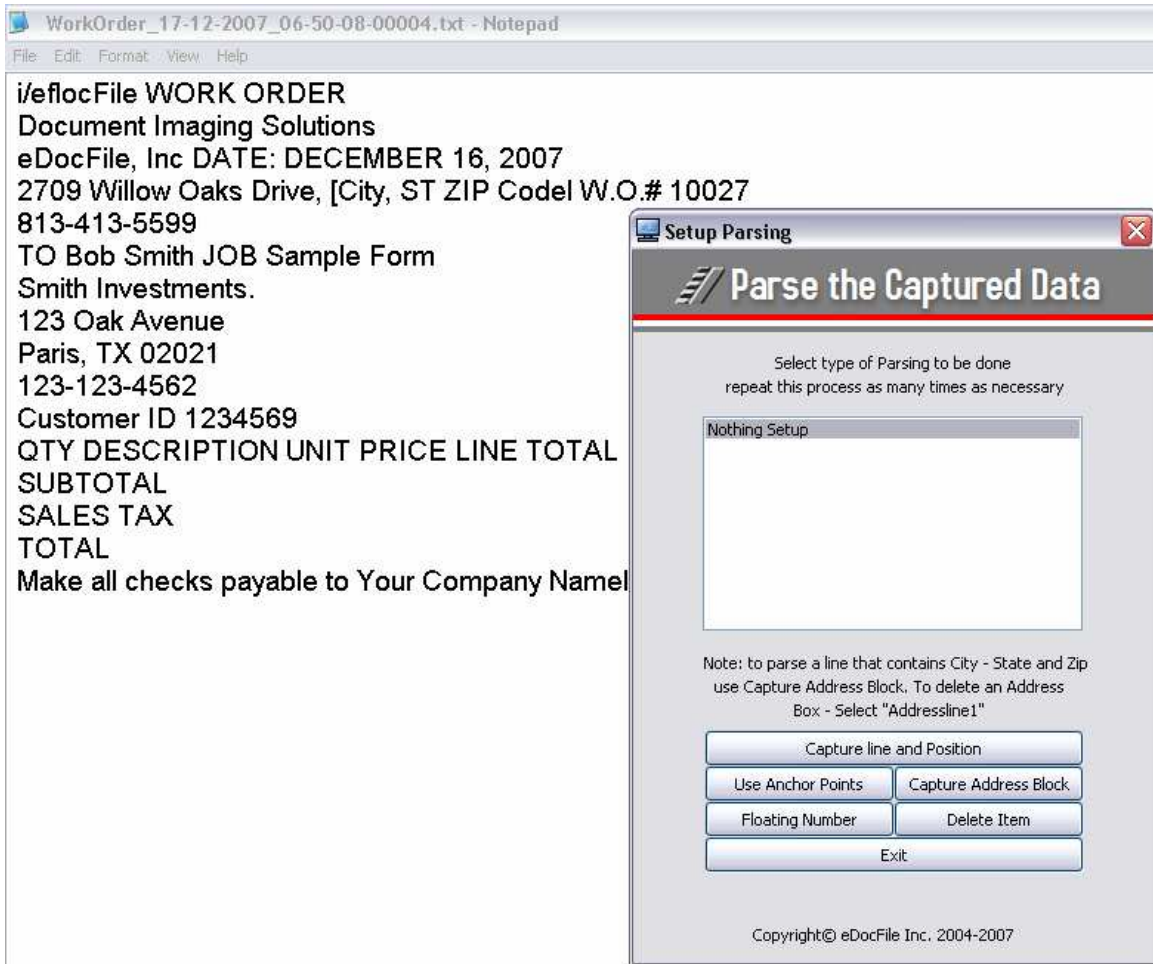


Enter a Job Name. The job name should match the document type as it can be used again for different output. For instance in the above examples for Work Orders and Receipts, everything would be consistent in the Work Orders except for the name of the person who performed the work, so this would only have to be setup once as later in the setup process the person's name can be entered in the output process. The same goes for receipts, the department can be added later.

Browse to a file that was scanned or a file in the output folder from the copier.

Click on Continue.





The selected file will be OCR'd, the text extracted and opened in the default text editor and a menu placed on top.

In this example the Work Order Number, the Customer ID, and the Clients name and address and phone will be captured. Later a folder will be setup for the customer with all of the Work Orders that correspond to them in it. The file name will be the Work Order and a csv file will be created that contains all the data

To capture the Work Order Number Anchor Points will be used. Anchor Points allow the user to collect data on a given number of lines after certain text appears. Click on Anchor Points.

A new Window will open

**Parse with Anchor Line**

## Anchor Point Line Parsing

Anchor Point Line Capture is for capturing text that changes depending upon the content of the Window

→  Name for this Value

Notes: on Naming Values  
Names: To parse a Full Name into First Name - Middle Initial - Last Name save the captured Value as FullName

\_\_\_\_\_ Anchor Information \_\_\_\_\_

Anchor Point One

Anchor Point Two

Lines to Add after Anchor Point

→  Line Containg Text

\_\_\_\_\_ Enter Start Position of Text to Capture \_\_\_\_\_

Static Position  or After Text  →

\_\_\_\_\_ Enter End Position of Text to Capture \_\_\_\_\_

Length of Capture  or Until Text

→  Check for End of Line

Copyright© eDocFile Inc. 2004-2007

The name WO\_numbl was entered as the name for the value, the program will look for the first occurrence of W.O.# in the text, once it finds it, the text after it will be captured until the end of the line.

Click on Test

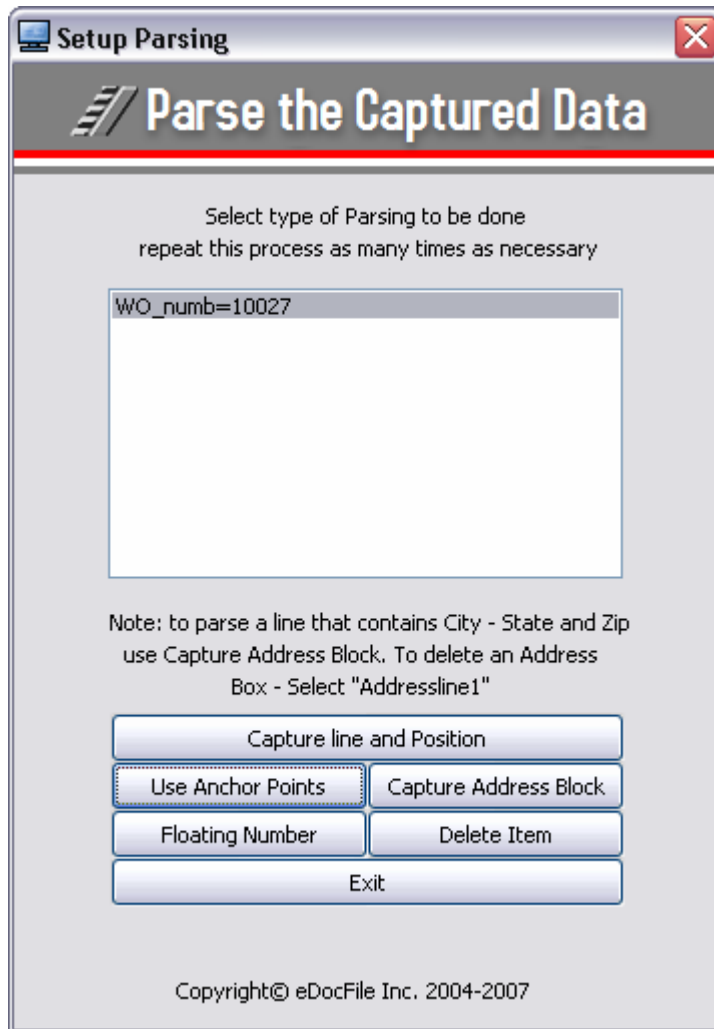
A Window shows the captured results.



Click on Okay, if it captured the data correctly click on Save and Exit, if not, try a different method.



Click on Yes to save the settings.



The captured data is displayed in the setup menu. Click on "Capture Address Block".

WorkOrder\_17-12-2007\_06-50-08-00004.txt - Notepad

File Edit Format View Help

i/flocFile WORK ORDER  
 Document Imaging Solutions  
 eDocFile, Inc DATE: DECEMBER 16, 2007  
 2709 Willow Oaks Drive, [City, ST ZIP Code] W.O.# 10027  
 813-413-5599  
 TO Bob Smith JOB Sample Form  
 Smith Investments.  
 123 Oak Avenue  
 Paris, TX 02021  
 123-123-4562  
 Customer ID 1234569  
 QTY DESCRIPTION UNIT PRICE LI  
 SUBTOTAL  
 SALES TAX  
 TOTAL  
 Make all checks payable to Your Cor

Auto Capture Address Fields

**Setup Address Fields Parsing**

Setup Parsing Address Field

This will Find three lines of an Address Field and create  
 Address1 - Address2 - City - State - Zip

Data to Exclude in Output

Anchor Point One Text  
 Sample Form

Anchor Point Two Text

Exclude Lines Containing  
 Cust

Test  
 Exit

Copyright© eDocFile Inc. 2004-2007

Capture Address Block looks for an address by finding a line that contains a two letter state abbreviation followed by a 5 digit zip code.

Sample Form is entered as the Anchor Point One, so the program will not start searching for the state until this is found. Exclude Lines Containing has Cust “shorted customer” in it so that it will not find ID thinking it is Idaho as it has a five digit number after it.

Click on Test

Address Box Parsing Message

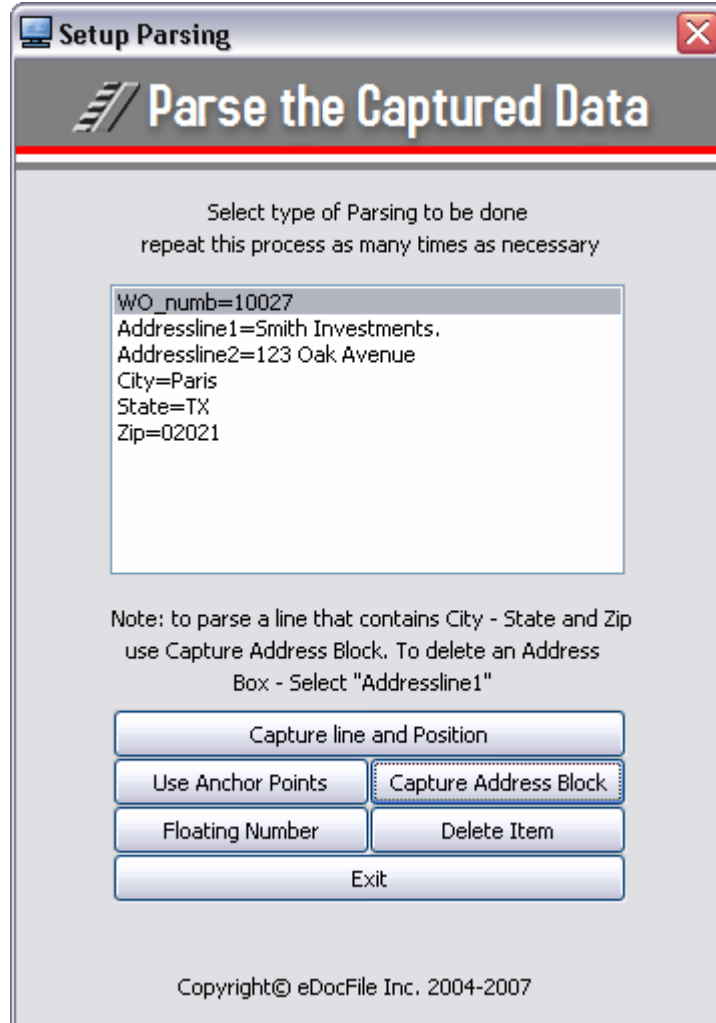
12/17/2007 3:19:56 PM

Smith Investments.  
 123 Oak Avenue  
 Paris  
 TX  
 02021

OK

The results will be shown.

Click on Exit and Yes to Save



The Address contents have been added. Click on "Floating Number" to capture the Customer ID.

WorkOrder\_17-12-2007\_06-50-08-00004.txt - Notepad

File Edit Format View Help

i/flocFile WORK ORDER  
Document Imaging Solutions  
eDocFile, Inc DATE: DECEMBER 16, 2007  
2709 Willow Oaks Drive, [City, ST ZIP Code] W.O.# 10027  
813-413-5599  
TO Bob Smith JOB Sample Form  
Smith Investments.  
123 Oak Avenue  
Paris, TX 02021  
123-123-4562  
Customer ID 1234569  
QTY DESCRIPTION UNIT  
SUBTOTAL  
SALES TAX  
TOTAL  
Make all checks payable to

Setup Floating Number

### Floating Number Setup

This function finds a Number that contains a specified number of Digits anywhere in the file

CustID  Name this Value

7 Specify number of Digits

1 Specify Occurance

Test Exit

Copyright© eDocFile Inc. 2004-2007

A new Window opens and it will search for an occurrence of number and the number of digits in the number. In this example, the customer number is the first occurrence of a number with 7 digits in it. This could also be captured by specifying the line the number is on and the text that comes after ID.

Click on Test

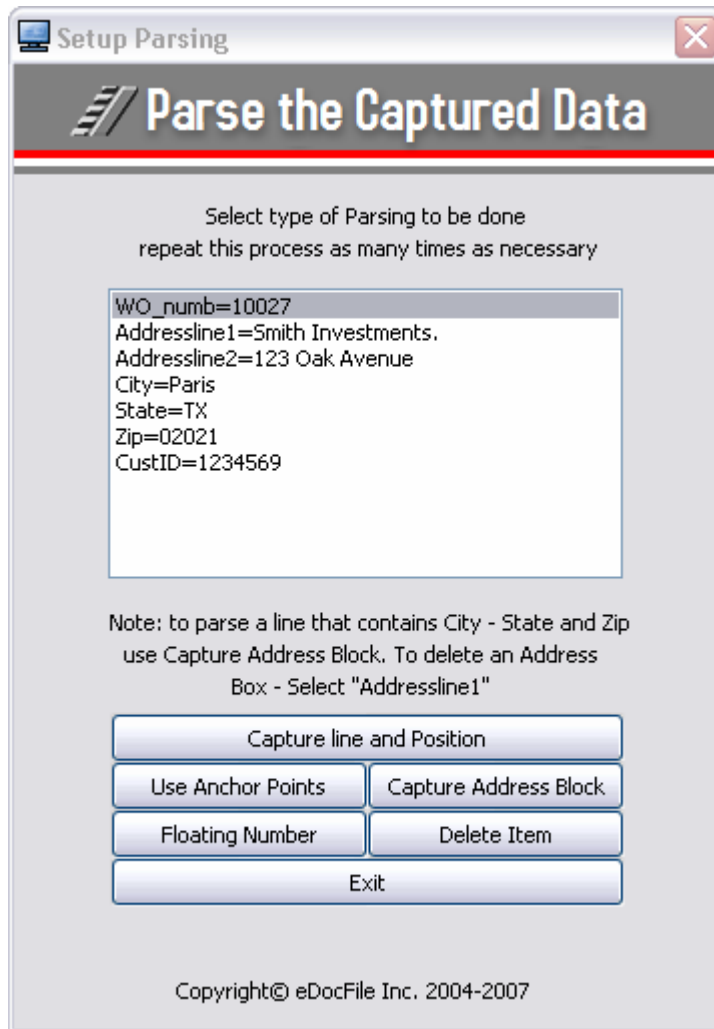
Floating Number Message

12/17/2007 4:31:41 PM

1234569

OK

The proper number is displayed. Click on OK, Exit and Yes to save the settings and return to the Setup Parsing Menu.



Click on Capture Line and Position to capture the phone number.



Parse Static Line

## Parse Static Line and Position

Static Line Capture is for capturing text that always appears on the same line each time the screen is captured

Name for this Value

Notes:  
Names: To parse a Full Name into First Name - Middle Initial - Last Name save the captured Value as FullName

\_\_\_\_\_ Line to Capture \_\_\_\_\_

Enter Starting Line Number


\_\_\_\_\_ Enter Start Position of Text to Capture \_\_\_\_\_

Static Position  or After Text

\_\_\_\_\_ Enter End Position of Text to Capture \_\_\_\_\_

Length of Capture  or Until Text

Check for End of Line




Copyright© eDocFile Inc. 2004-2007


Enter phone for the name of the value and click on Setup Wizard.

A new Window will open, Select the phone number by clicking on it and dragging the mouse over the number to highlight it. Then click on Okay.

(If multiple Windows are open and the focus goes to Notepad, press ALT + Tab to return to the menu.)

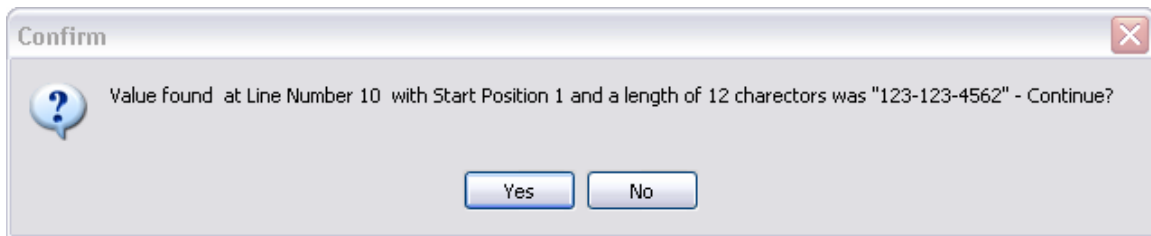
The Captured Value will be displayed.

**Confirm** 

 Is 123-123-4562 the correct value?

Click on Yes to Continue, no to try again

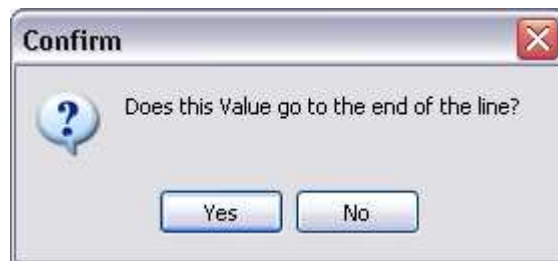
A confirmation menu appears showing the line and position.



Click on Yes to continue or No to repeat the process



A window prompts the user for data that always precedes the value to be captured, in this case the answer is no.

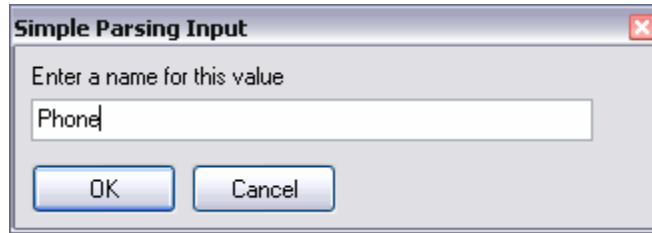


A Window prompts to see if the data goes to the end of the line. In this case the answer is Yes.

Notes: it text appears before the value to be captured the user will be prompted to enter it, the same if text appears after it. So if the line contained  
Phone: 813-813-4562 ext:12

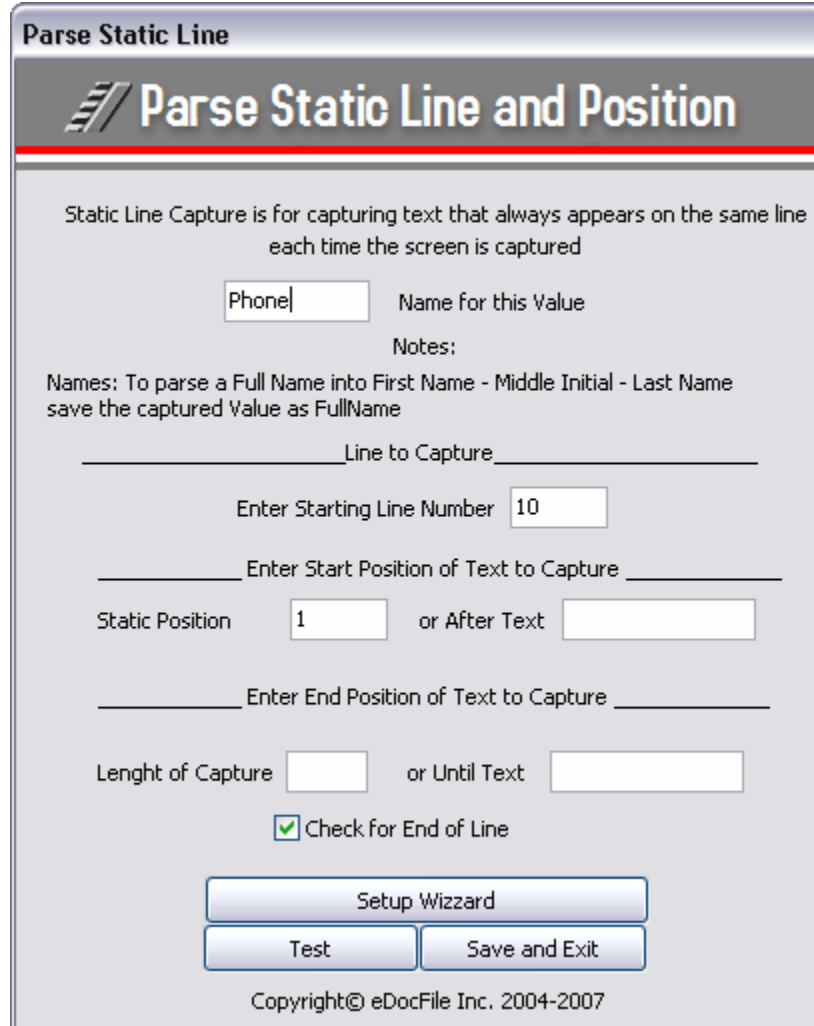
The user could enter Phone: and then ext: and the number between them would be captured. Notice how the 12 was not entered as it will not always be 12 it could be a different extension number.

Enter a name for the value after this is complete.



A small dialog box titled "Simple Parsing Input" with a close button in the top right corner. It contains a text input field with the text "Phone" entered. Below the input field are two buttons: "OK" and "Cancel".

Click on OK



A larger dialog box titled "Parse Static Line" with a decorative header bar containing the text "Parse Static Line and Position". Below the header, there is a paragraph of text: "Static Line Capture is for capturing text that always appears on the same line each time the screen is captured".

Below this text is a text input field containing "Phone" and the label "Name for this Value".

Underneath is the text "Notes:" followed by "Names: To parse a Full Name into First Name - Middle Initial - Last Name save the captured Value as FullName".

There are several input fields and labels:

- "Line to Capture" with a blank input field.
- "Enter Starting Line Number" with an input field containing "10".
- "Enter Start Position of Text to Capture" with a blank input field.
- "Static Position" with an input field containing "1" and "or After Text" with a blank input field.
- "Enter End Position of Text to Capture" with a blank input field.
- "Length of Capture" with a blank input field and "or Until Text" with a blank input field.
- A checked checkbox labeled "Check for End of Line".

At the bottom, there are three buttons: "Setup Wizard", "Test", and "Save and Exit".

At the very bottom, it says "Copyright© eDocFile Inc. 2004-2007".

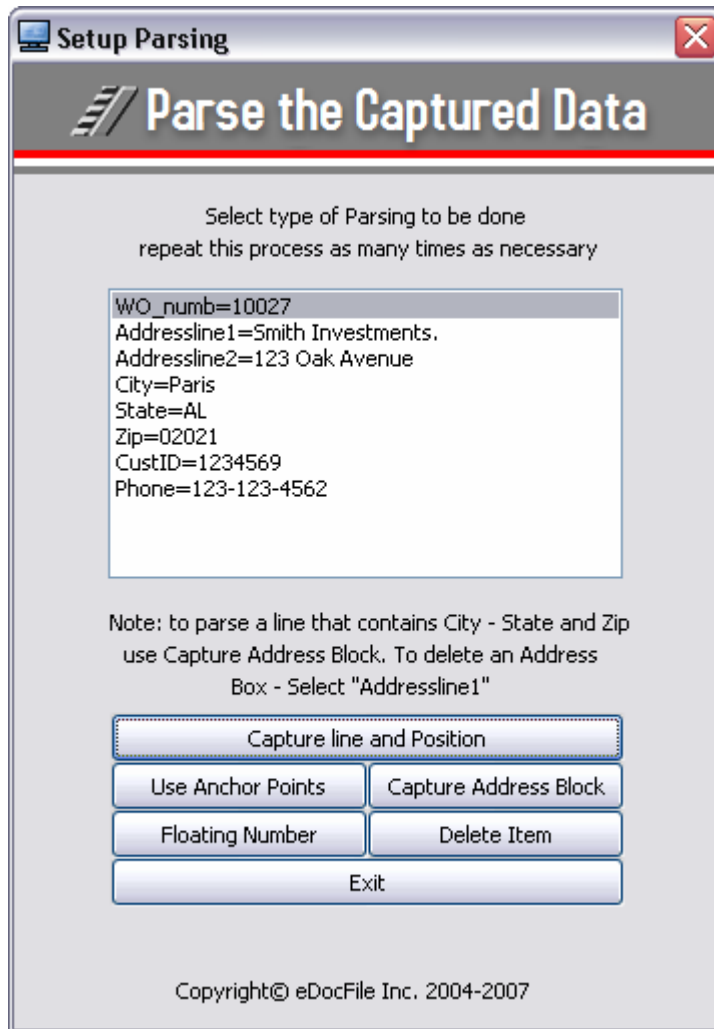
Click on Test to Check the parsing.



Click on OK, then on Save and Exit



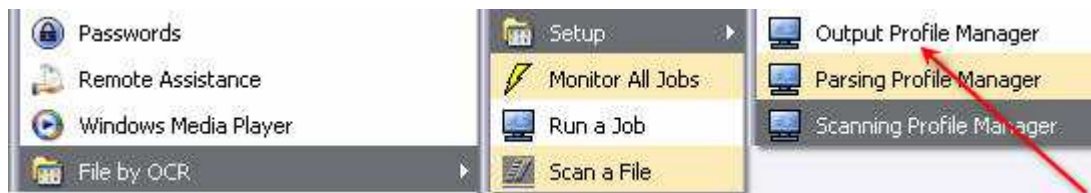
Click on Yes



Check the Captured data if it is all correct click on Exit, if not highlight the line, delete the item and parse it using a different method.

### ***Step Three – Generating Output***

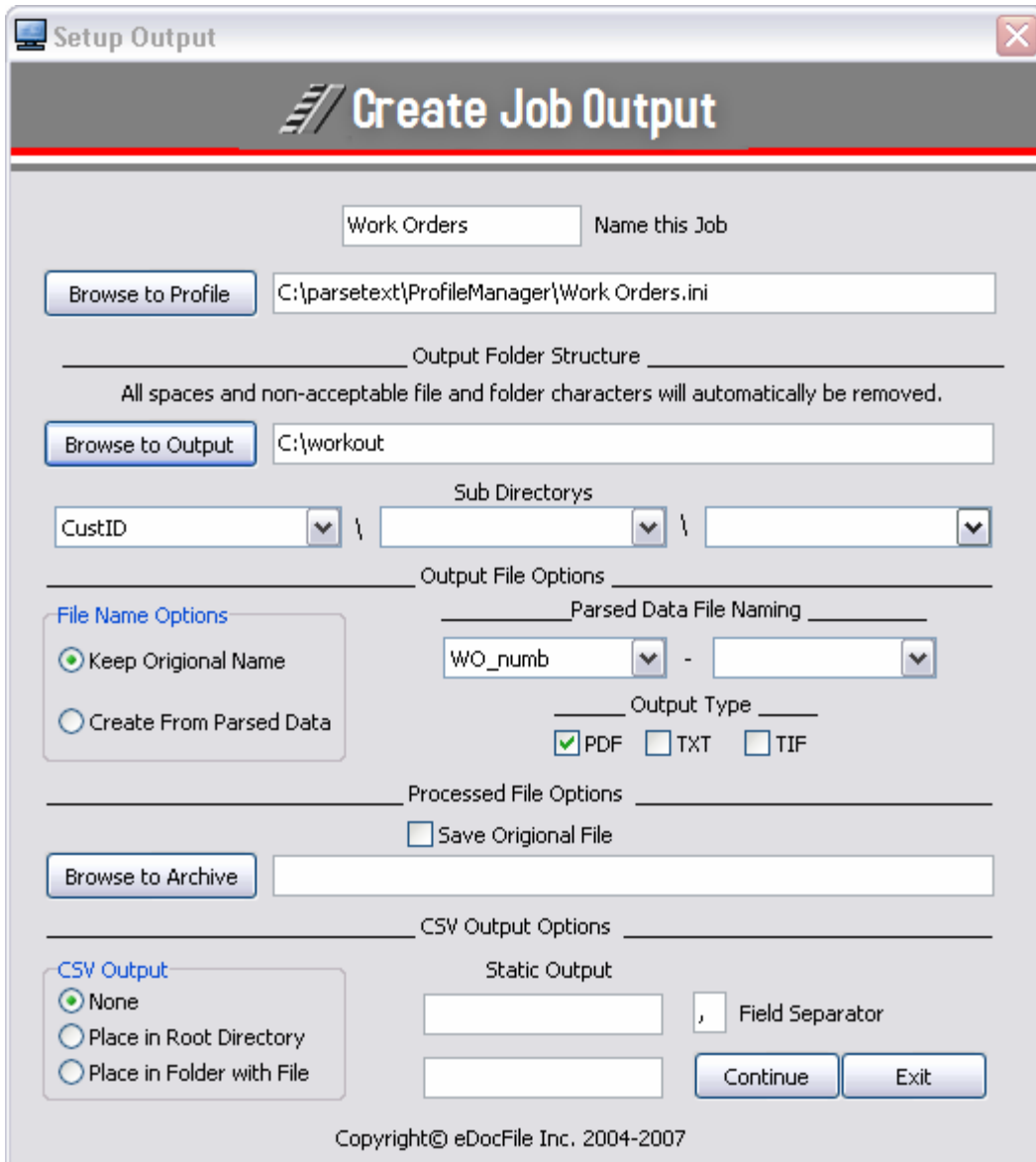
To begin click on Start/All Programs/File by OCR/Setup/Output Profile Manager.



Click on New



A new Window will open



Enter a name for this output and select a profile using the Browse to Profile button. This will select the parsing job that was completed in step two.

Create a root directory for the output, either type in the path or if it already exists select it with the Browse to Output button.

Select Sub Directories to be created from the parsed data. In this case CustID is being used which contains the customer number. So the output will have a root directory, and sub directories for each customer. The customer name could have been selected here as well. In the Output File Options check Create from Parsed Data, as the file name will be created, if keep original Name was selected and the documents were entered from a scanner the Keep Original Name option would be useful if the documents were scanned in a batch that was separated by employee. This would make the file name the employees

name and date of capture. With this if all other parsed data was incorrect the document could still be found.

From the drop down list WO\_num was selected for Work Order Number so the structure will be:

Root Directory\Customer Number\Work Order Number.

Click on Save Original and browse to a path for original images if they are to be saved.

Under CSV Output Options select place in Root Directory. This will create a delimited file with all captured data. This data can later be searched or imported into a data base for searching. The option is also available to place this is the output folder with the file. Static output can also be added. In the example above with the work orders being scanned in a batch by employee, the employee's name can be entered here.

Select a delimiter, the most common are comma, and semicolon but anything can be entered.

Once finished click on Continue to Exit.

## Step Four - Running the Job

There are two options for running the job, one is to run a single job manually and the other is to monitor all file folders and process each document as it comes in. OCR is very processor intensive and this should be considered when making the selection.

To Run a single Job

To begin click on Start/All Programs/File by OCR/Run a Job



A new window will open





Select the Job to Run and click on Run Job

Or Start/All Programs/File by OCR/Monitor All Jobs



If this option is selected all folder will be monitored and processed. An icon will appear in the Sys Tray to Start / Stop or exit the process.

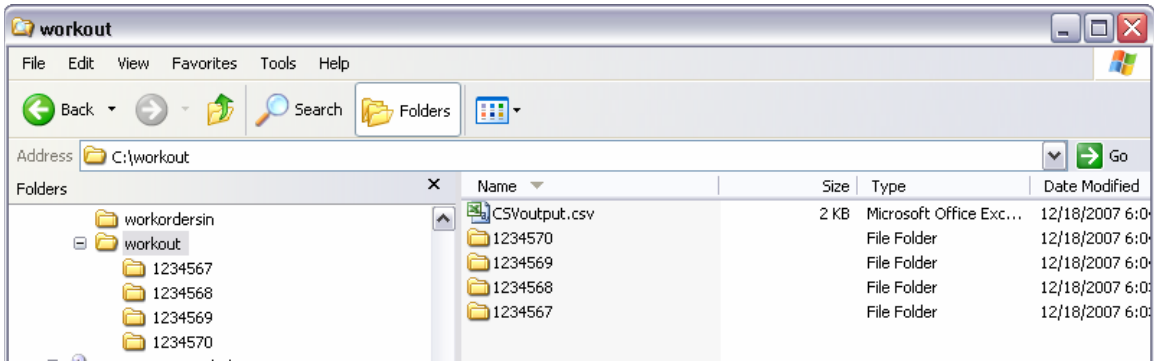


To access these options right click on the Icon.



Step Five – Check the output

Open explorer and open the folder containing the output.

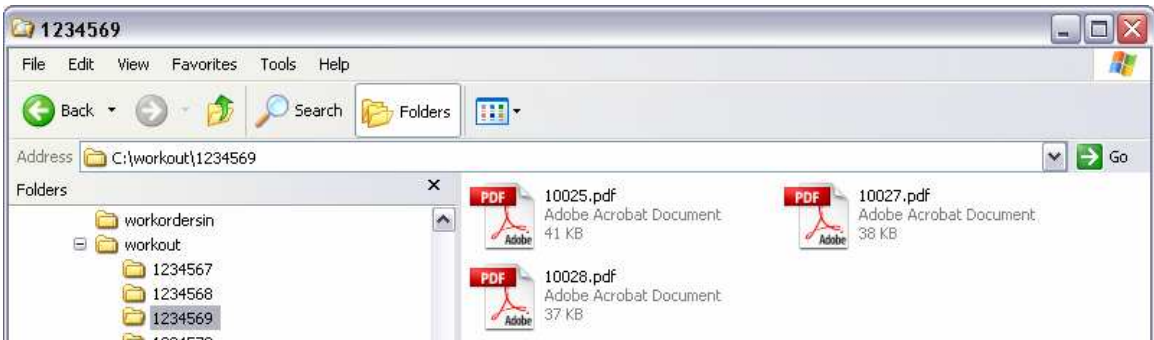


It contains a file folder for each client and a csv file

Static Output	Static Output 2	File Processed	PDF Output	VO_num	Addressline1	Addressline2	City	State	Zip	CustID	Phone
		C:\parsetest\temp\images\workorders_18-12-2007_06-02-55-000001.tif	C:\workout\1234569\10028.pdf	10028	Smith Investments.	123 Oak Avenue	Paris	TX	2021	1234569	123-123-4562
		C:\parsetest\temp\images\workorders_18-12-2007_06-02-55-00005.tif	C:\workout\1234569\10025.pdf	10025	Smith Investments.	123 Oak Avenue	Paris	TX	2021	1234569	123-123-4562
		C:\parsetest\temp\images\workorders_18-12-2007_06-02-55-00004.tif	C:\workout\1234569\10023.pdf	10023	Consolidated Corp.	123 Main Street	Paris	TX	2021	1234567	123-123-2313
		C:\parsetest\temp\images\workorders_18-12-2007_06-02-55-00003.tif	C:\workout\1234569\10024.pdf	10024	Acme Corp.	123 Maple Avenue	Paris	TX	2021	1234568	123-123-0000
		C:\parsetest\temp\images\workorders_18-12-2007_06-02-55-00002.tif	C:\workout\1234569\10027.pdf	10027	Smith Investments.	123 Oak Avenue	Paris	TX	2021	1234569	123-123-4562
		C:\parsetest\temp\images\workorders_18-12-2007_06-02-55-00001.tif	C:\workout\1234570\10026.pdf	10026	Ace Plumbing Corp.	123 Oak Avenue	Paris	TX	2021	1234570	123-123-4561

A quick review will show any obvious mistakes in the data capture and parsing. In this example everything was captured correctly except for the spacing on two phone numbers.

When opening the file folder



The file names are the same as the work order number and are completely text searchable.

## Trouble Shooting

The document used in this example was designed poorly for data capture with OCR.



# WORK ORDER

eDocFile, Inc  
2709 Willow Oaks Drive  
813-413-5599

DATE: DECEMBER 18, 2007  
→ W.O.# 10028

TO Bob Smith  
Smith Investments.  
123 Oak Avenue  
Paris, TX 02021  
123-123-4562  
Customer ID 1234569 ←

JOB Sample Form

QTY	DESCRIPTION	UNIT PRICE	LINE TOTAL

The Work Order number is a standard size 12 point font as is the Customer ID.

Ideally the form would be redesigned with larger OCR fonts, which would improve accuracy. If using a somewhat standard font make it very standard, for instance instead of using the font Garland choose Times New Roman or Arial.

ID 36923

JOB 654987

# WORK ORDER

DATE: DECEMBER 20, 2007

TO  
A.B. Dick Products Co of Rockford  
5027 Harrison Avenue  
Rockford, IL 61108  
Phone: 815/397-0660

JOB Sample Form

QTY	DESCRIPTION	UNIT PRICE	LINE TOTAL
-----	-------------	------------	------------

The example above would be ideally suited to for the capture and parsing of the data.  
Test with 50 documents it captured all characters successfully.