DocScan

Installation guide and user manual

February 2013

Knowledgeone corporation

Copyright notice

Every effort has been made to ensure that the information in this document is up to date and accurate. Knowledgeone Corporation welcomes advice of any changes or corrections for the next edition.

Enquiries

E-mail

To easily obtain information by e-mail, send enquiries to:

Sales	sales@knowledgeonecorp.com
Support	support@knowledgeonecorp.com
Training	training@knowledgeonecorp.com

Technical support

For technical support questions or requests, we encourage you to contact our International Support Center.

E-mail	support@knowledgeonecorp.com
Toll-free	United States — 1888 325 1614 Canada — 1888 405 9019 United Kingdom — 0808 234 8828 Australia — 1800 221 061 (excluding Sydney) New Zealand — 0800 445 438 (Sydney customers please phone 8913 9300.)

To speed the technical support process, please note the following before contacting the International Support Center:

- K1 Corp Customer Number
- K1 Corp Incident PIN
- Product Version
- Type of Database Server (Oracle/MSSQL Server) and Version

Knowledgeone Corporation website

Visit our website at http://www.knowledgeonecorp.com/ for information on the latest K1 Corp products, support issues and training dates.

Table of Contents

Welcome to DocScan

DocScan 3.4, February 2013

DocScan works with the latest TWAIN specifications. It has been tested with a range of scanners to ensure that it is 100% compatible with the latest implementation of the TWAIN standard.

Please note, however, that DocScan is not a "general" scanning solution — rather, it has been specifically designed to capture multi-page text documents and store them in a format suitable for OCR applications. It is also designed to produce documents for RecScan and Xchange to process and "attach" to the RecFind relational database.

It contains code to recognize the identifying barcode (Code 3 of 9 only) on the first page of each document. DocScan then assumes that all pages (without a barcode) following a page with a barcode are subsequent pages of a single document. It stores all pages in a multi-page TIFF image using the barcode as the identifier.

The barcode used must be in Code 3 of 9 format as shown in the following example. The barcode must contain an asterisk (*) as both the first and last characters. See the following example displayed in both Code 3 of 9 and Arial format:

DocScan 3.4 works with both text images and graphical images in either black and white, grey-scale, or colour. Because of the need to search every scanned page for a barcode, DocScan will always run the scanner at a speed somewhat less than its optimum rating; this is normal because of the extra processing involved with each and every page.

What's New In This Version

Following are new features included in this version of DocScan:

Save As PDF/A-1

DocScan 3.4 gives you the ability to create PDFs that meet the PDF/A-1 standard. The PDFs will be created with Level B conformance (PDF/A-1b) to the standard which is sufficient for scanned documents even if they are processed using OCR to make them text-searchable.

Note: PDF/A-1b (Level B Conformance) meets the minimum requirements of the PDF/A-1 standard.

Import Using Xchange
DocScan 3.4 expands forms processing by adding an option "Import Using Xchange". If this option is selected DocScan will produce an XML file that can be used as a data source for Xchange to import into the Recfind database. This XML file will contain all information retrieve during OCR as well as an encoded version of the PDF. Using Xchange for import gives you the ability to add your information to any table (including custom tables) in the RecFind database.

Requirements

DocScan 3.4 requires that you have Microsoft's .NET 3.5 Framework (or higher) installed. If you received DocScan on a CD-ROM, the .NET framework would have been supplied to you. If you downloaded DocScan, the .NET framework can be obtained directly from Microsoft, at: http://www.microsoft.com/net/.

To use the OCR, convert-to-PDF functionality, or Forms processing in DocScan 3.4, you will need to have Microsoft Office Document Imaging (MODI), version 11 or higher, installed. MODI is available with Microsoft Office 2003 and 2007.

If you are using Microsoft Office 2003, MODI should be installed by default, and no further action will be required. If you are using Microsoft Office 2007, you will need to manually install this application before proceeding with your DocScan installation. We recommend that you obtain your Microsoft Office 2007 installation disc and install this program, as per Microsoft's instructions, first.

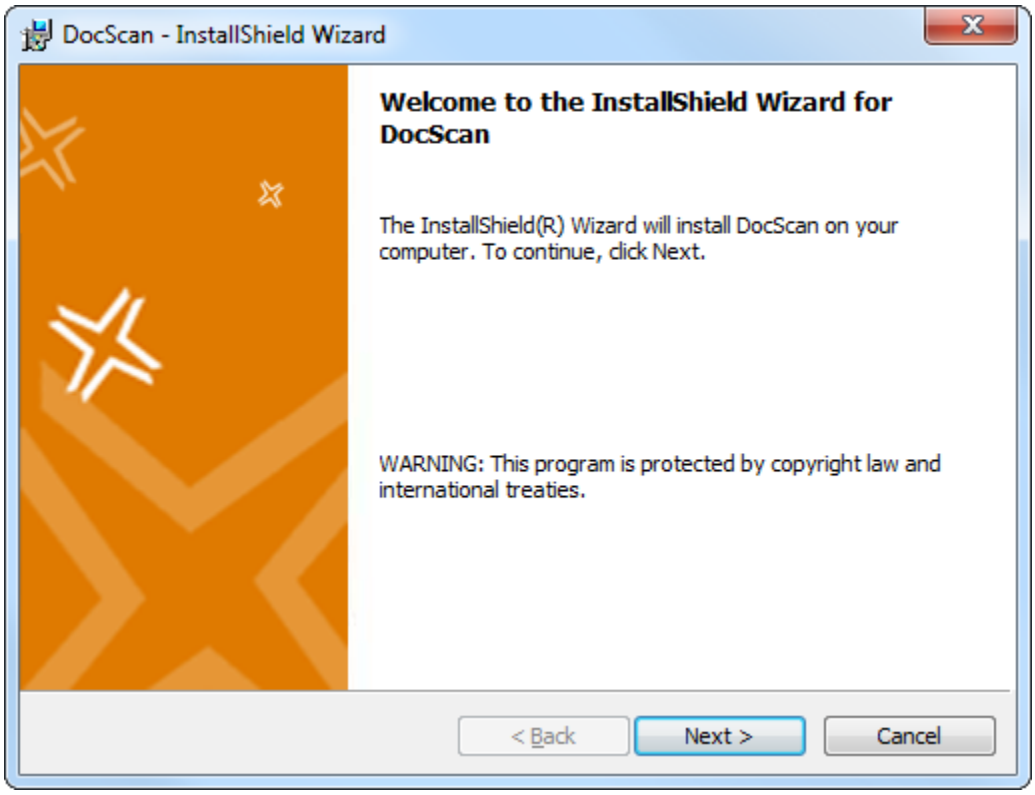
Issue with Microsoft Windows Server 2003
If you are using DocScan on a machine running Microsoft Windows Server 2003 (Service Pack 1), you may experience some problems with the OCR component of DocScan. This is a problem that Microsoft has acknowledged, and is described in detail at: http://support.microsoft.com/kb/918215/en-us/. For information on fixing this issue (directly from Microsoft), please see this link: http://support.microsoft.com/kb/875352/.

Installation and Set-up Guide

Installing DocScan

1: If you received DocScan on a CD-ROM, insert the disc and run the file "setup.exe" file located in the root directory. If you downloaded DocScan, you will have a "docscansetup.exe" file — run this.

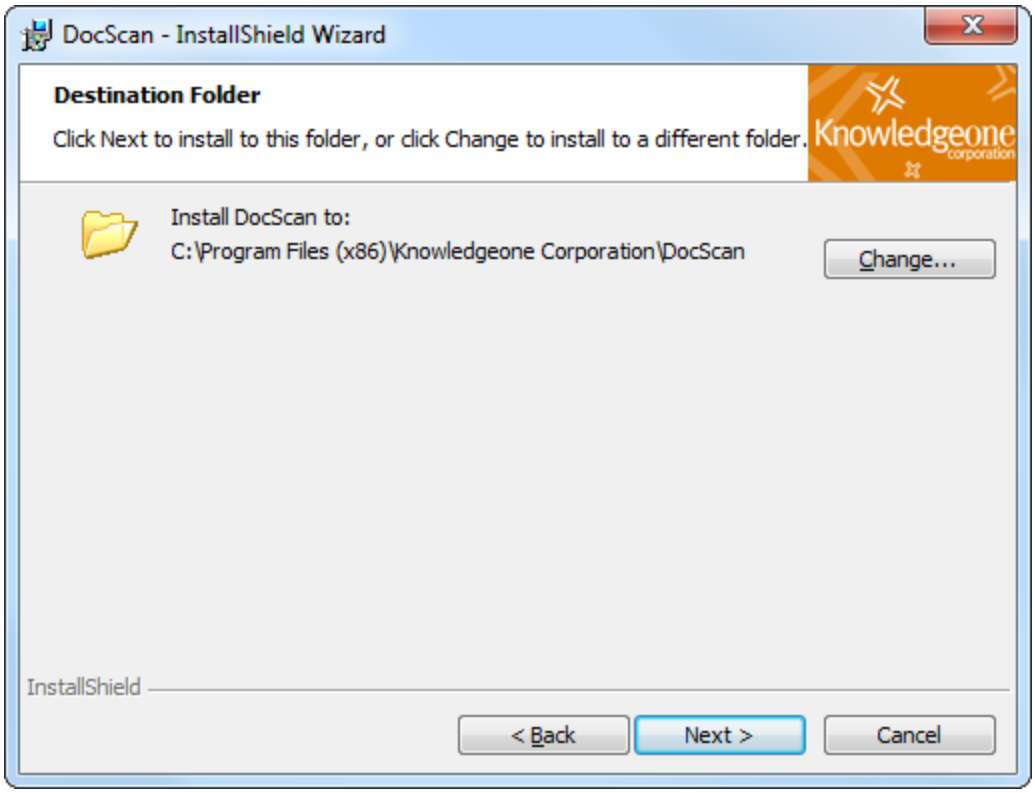
You will be presented with the following dialogue window:



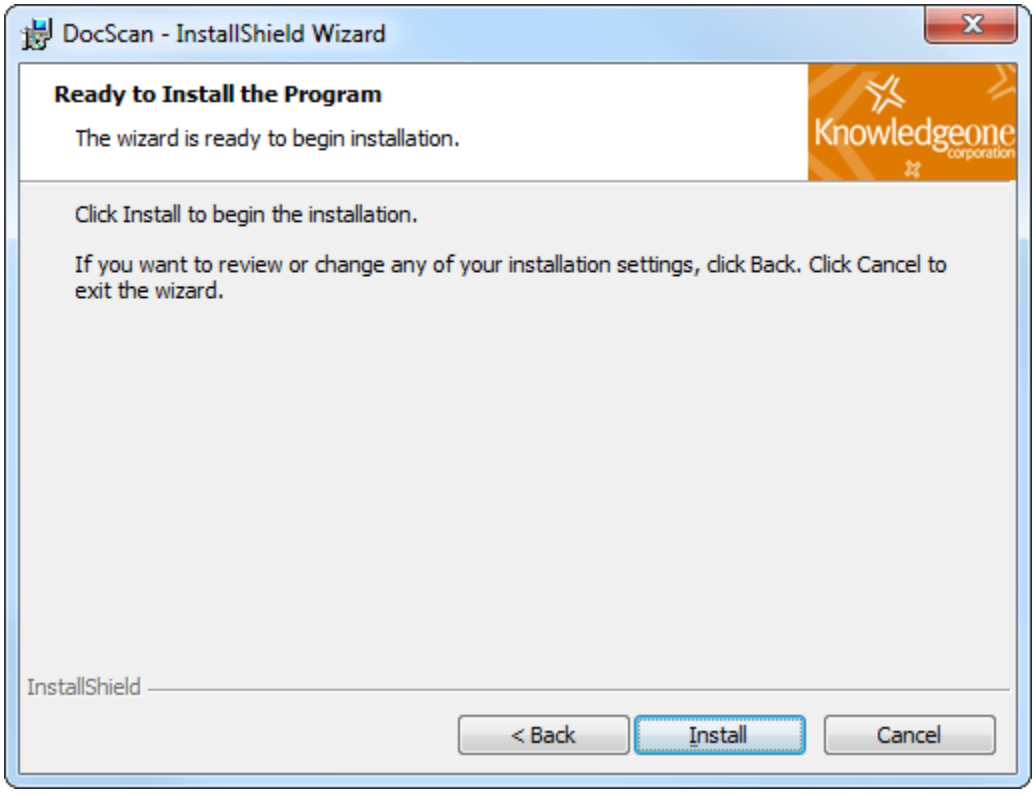
2: Click Next to continue.

3: Select "I accept the terms in the license agreement", and then click Next.

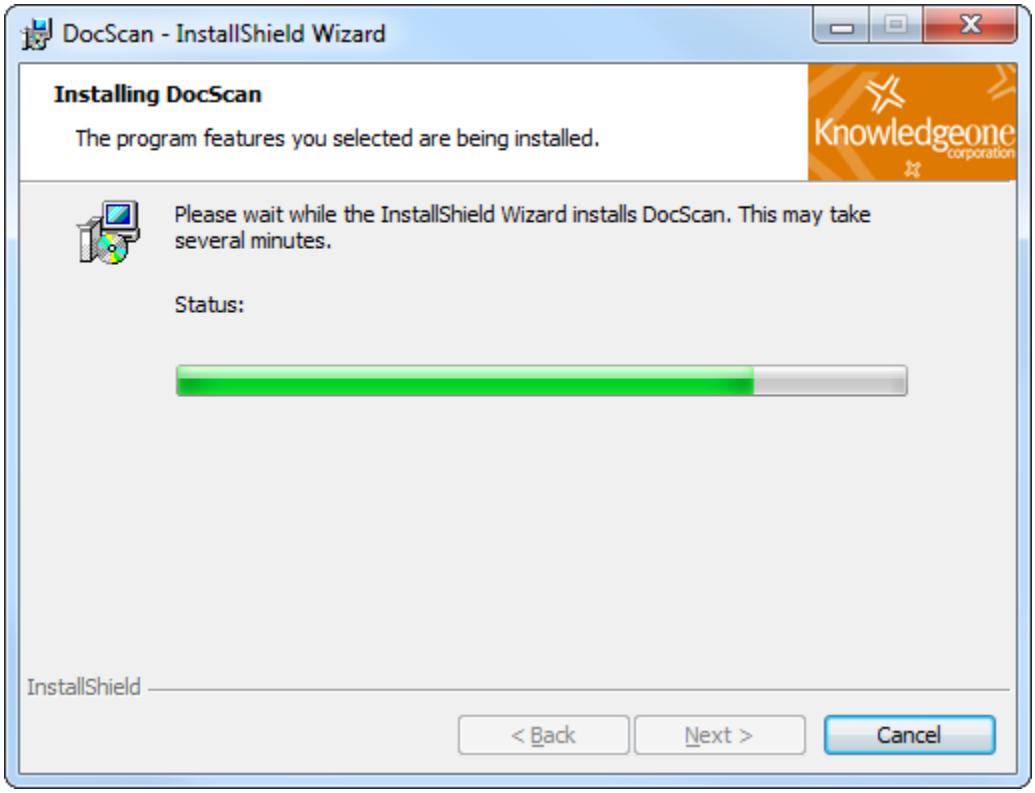
4: Select the destination folder for your DocScan installation. We recommend using the default setting; if you want to install the application to a different location, however, click the "Change" button to do so. Click Next to continue.



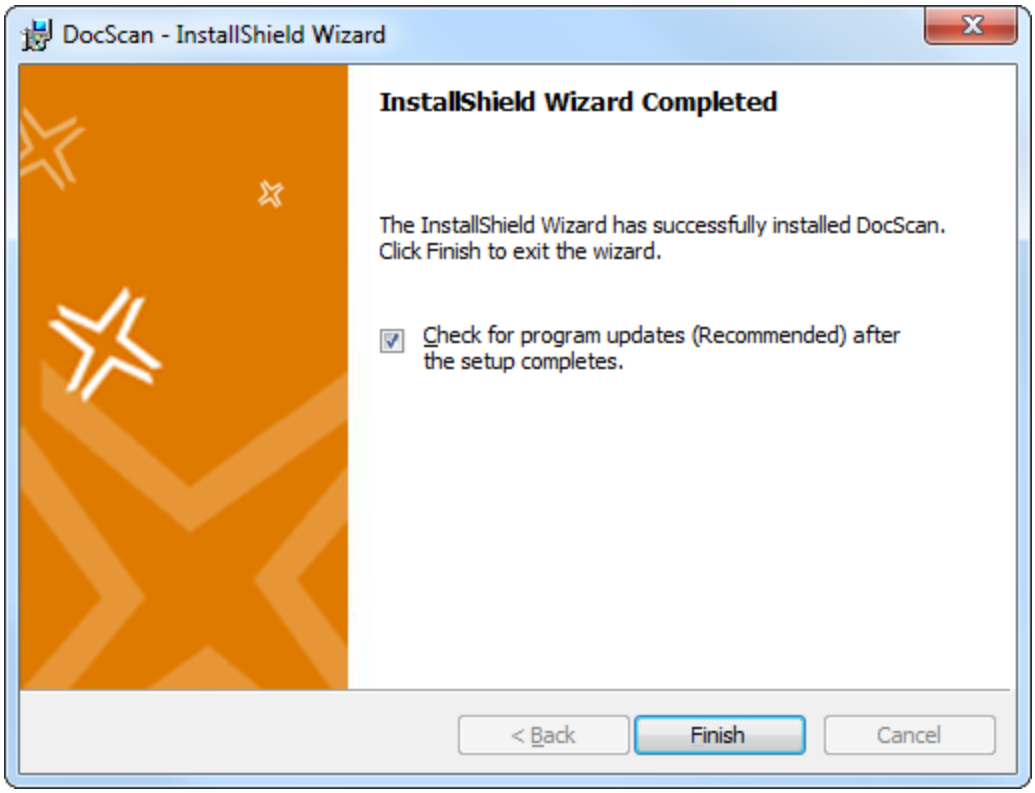
5: Click Install to begin the installation process.



6: You will be presented with the following dialogue window during the installation:



7: Click Finish to exit the installation wizard:



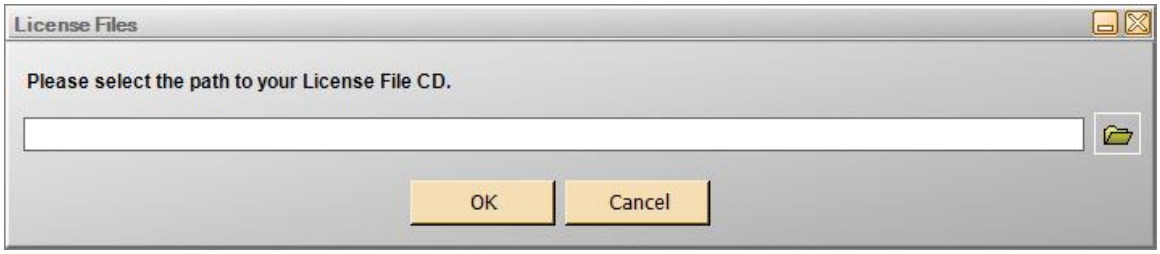
Registering DocScan

1: Once you have installed DocScan, you will need to register the program. If you do not register, DocScan will function for 45 days in a trial mode; all of the program's functionality will be present, but a watermark will be superimposed on scanned images and any generated PDF files.

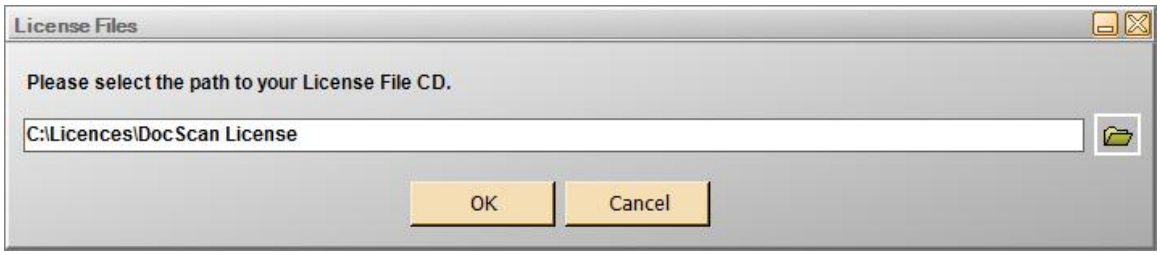
2: If you decide to register DocScan, you will need to obtain a license file from Knowledgeone Corporation. You can obtain this by contacting the company at **support@knowledgeonecorp.com**. The license file will be delivered to you on a CD-ROM.

3: Once you have the license file, you may register the program. To do this, open DocScan, and access the "Register" option on the "Help" pull-down menu.

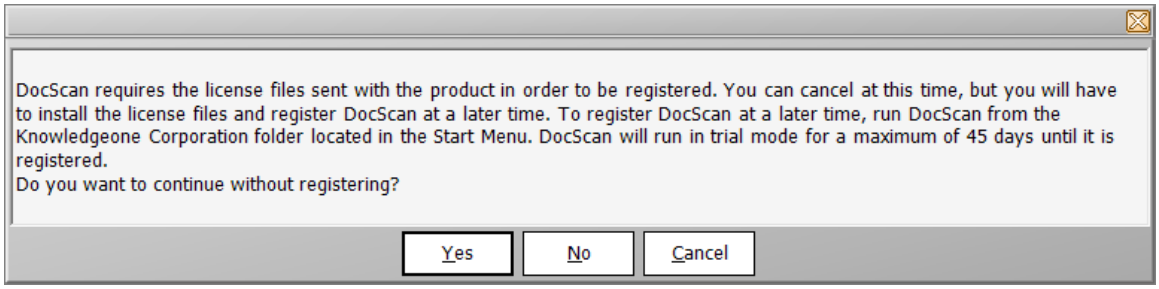
4: You will be asked for the location of your License File CD-ROM. If you are sure of the path, you can type it in directly. If you are unsure, click the "folder" icon to browse to the location using Windows Explorer.



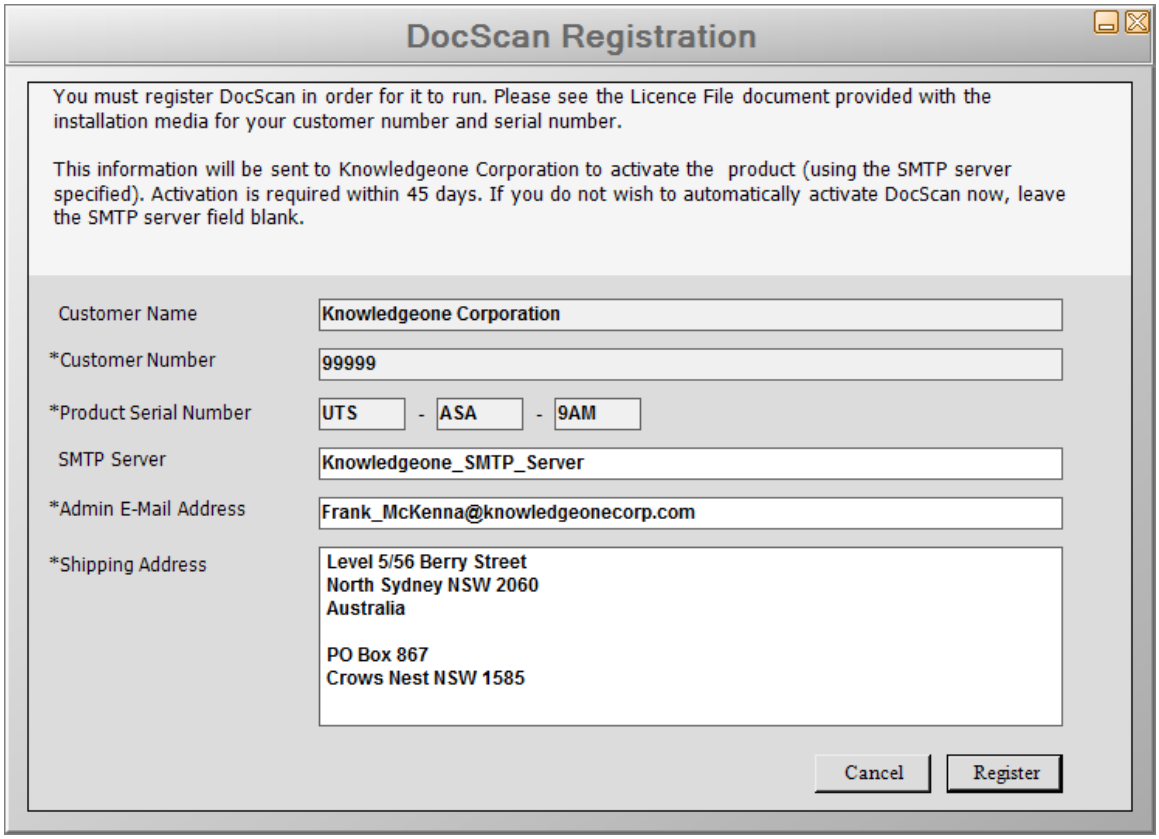
5: Once the path has been entered, click OK to continue.



If you click "Cancel" on Step 2, DocScan will inform you that it will not be able to properly run until you specify a valid license file:



6: You will then be presented with the DocScan registration screen.

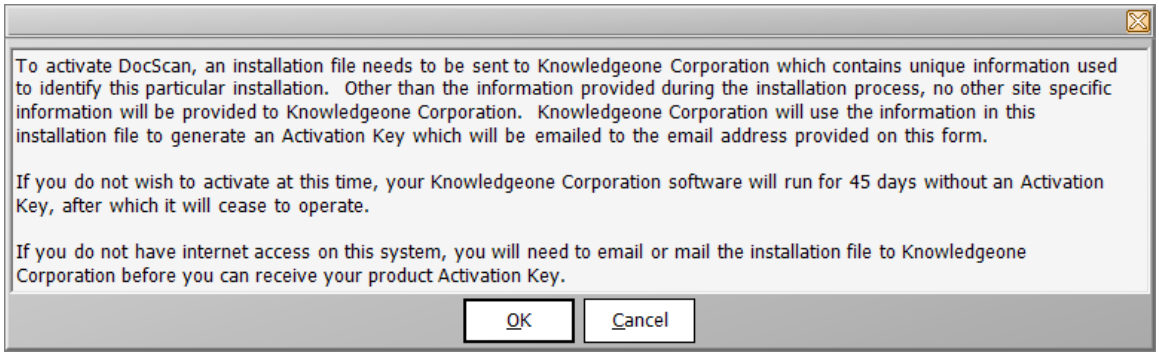


7: The customer name, number and product serial number will be automatically completed using the information in your licence file. Please enter your SMTP server (*this is your own mail server*), administrator's email address and you shipping address.

Note: An SMTP server is required to send this information to Knowledgeone Corporation automatically.

8: Once all of your information is entered click the Register button.

9: If you provided an SMTP server and e-mail address, you will be asked to confirm that you wish to register with Knowledgeone Corporation. If you do not register, DocScan will run for 45 days, after which time it will cease to operate.

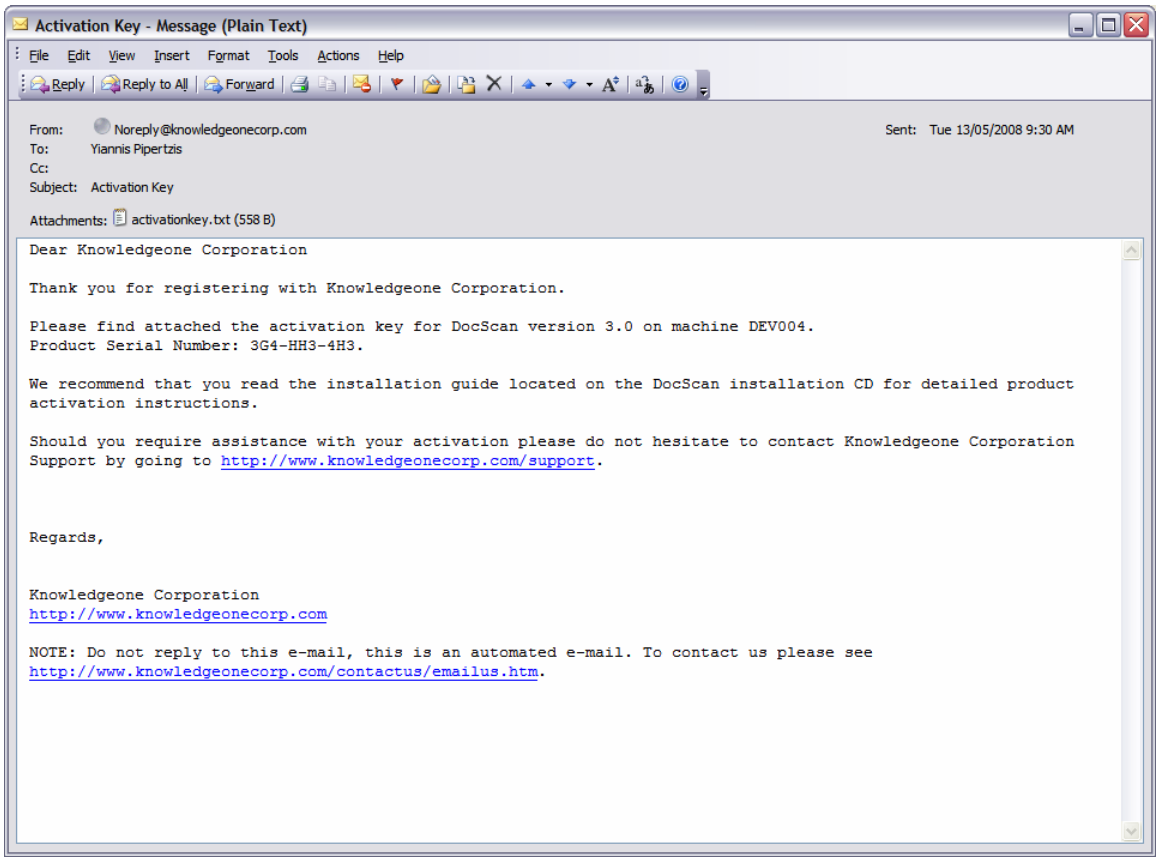


10: If you agree, click OK.

11: If the registration completes successfully, you will see the dialogue window shown below. Click OK.

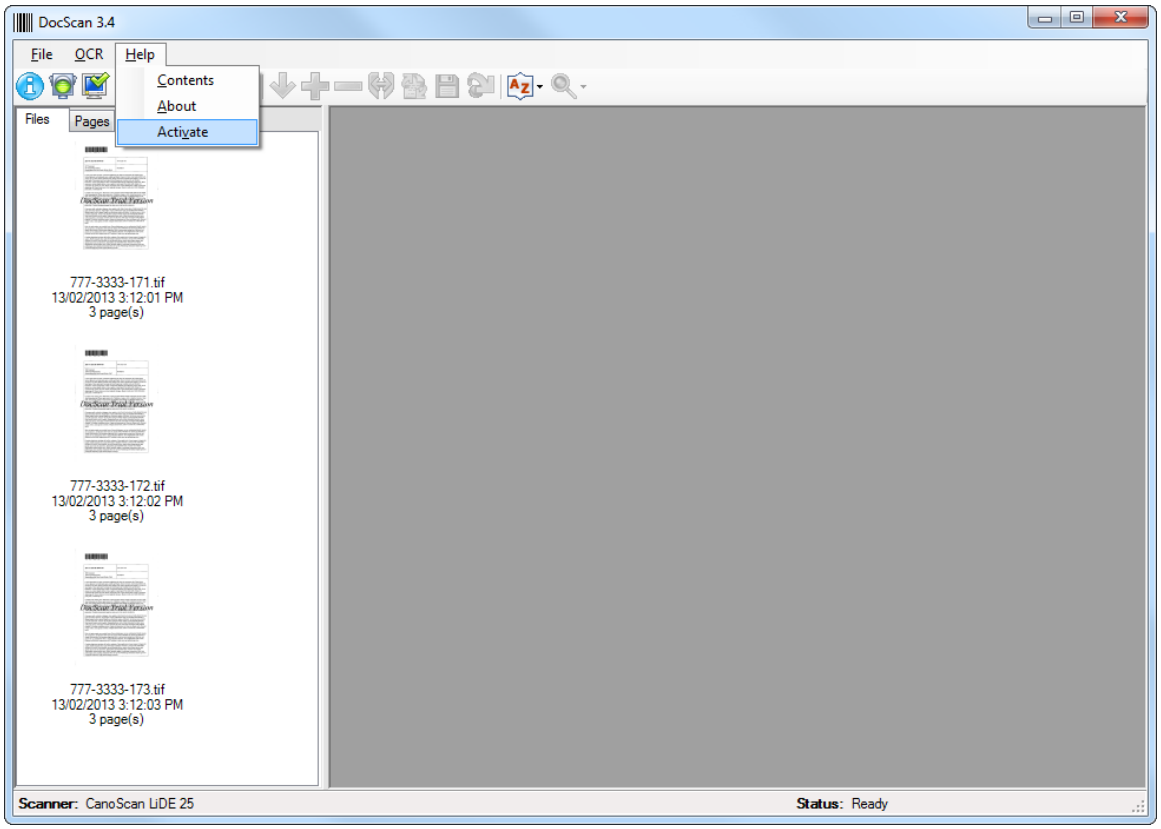
Activating DocScan

1: If you registered your installation with Knowledgeone Corporation, you should receive an e-mail providing you with your activation key. It will appear similar to the e-mail presented below:

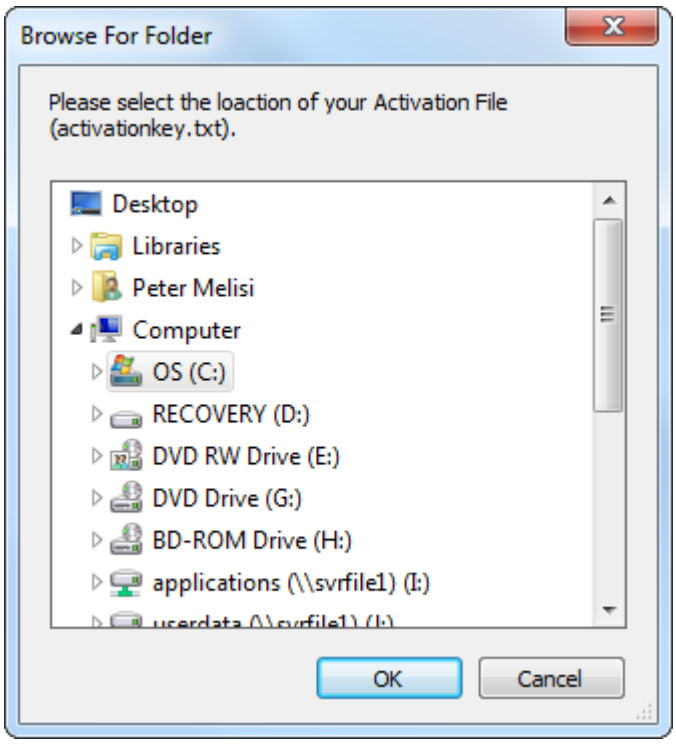


2: Copy the attached “activationkey.txt” file to a location on your hard drive. It is important that you remember where this is.

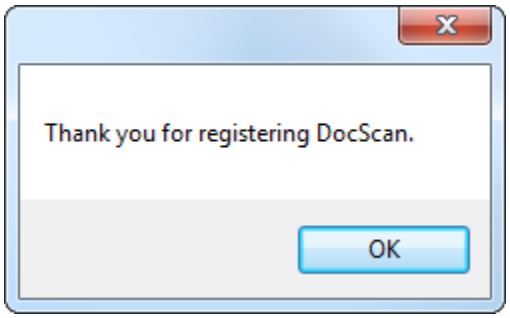
3: Run DocScan, and click the “Activate” option on the “Help” pull-down menu (as shown in the screenshot below).



4: DocScan will then ask you to specify the location of your Activation Key. Supply the path to which you saved the file in Step 2, and click OK:

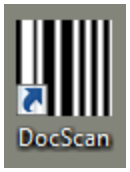


5: DocScan will then provide the following dialogue box. Congratulations, your installation of DocScan is now complete.

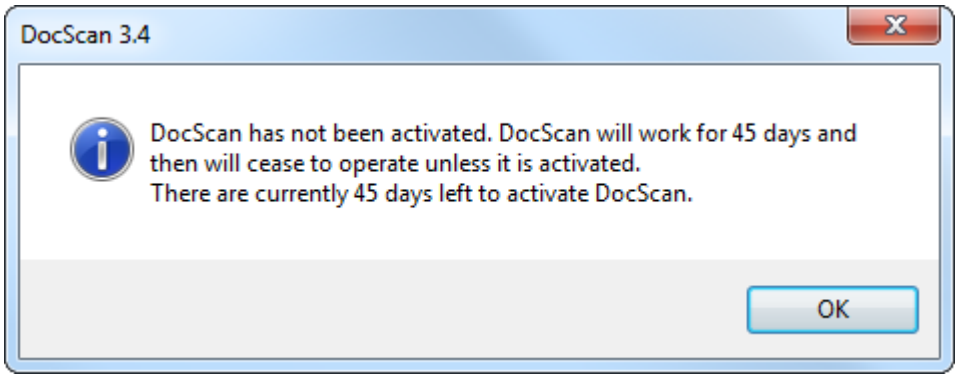


Getting Started

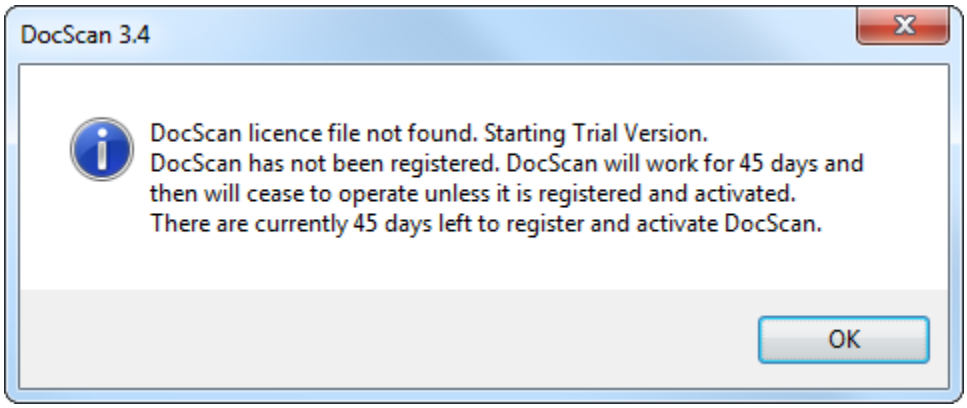
To start DocScan, double-click on the DocScan icon on your desktop.



If you have not activated your copy of DocScan a screen will appear stating the number of days left to activate your DocScan 3.4. Click OK.

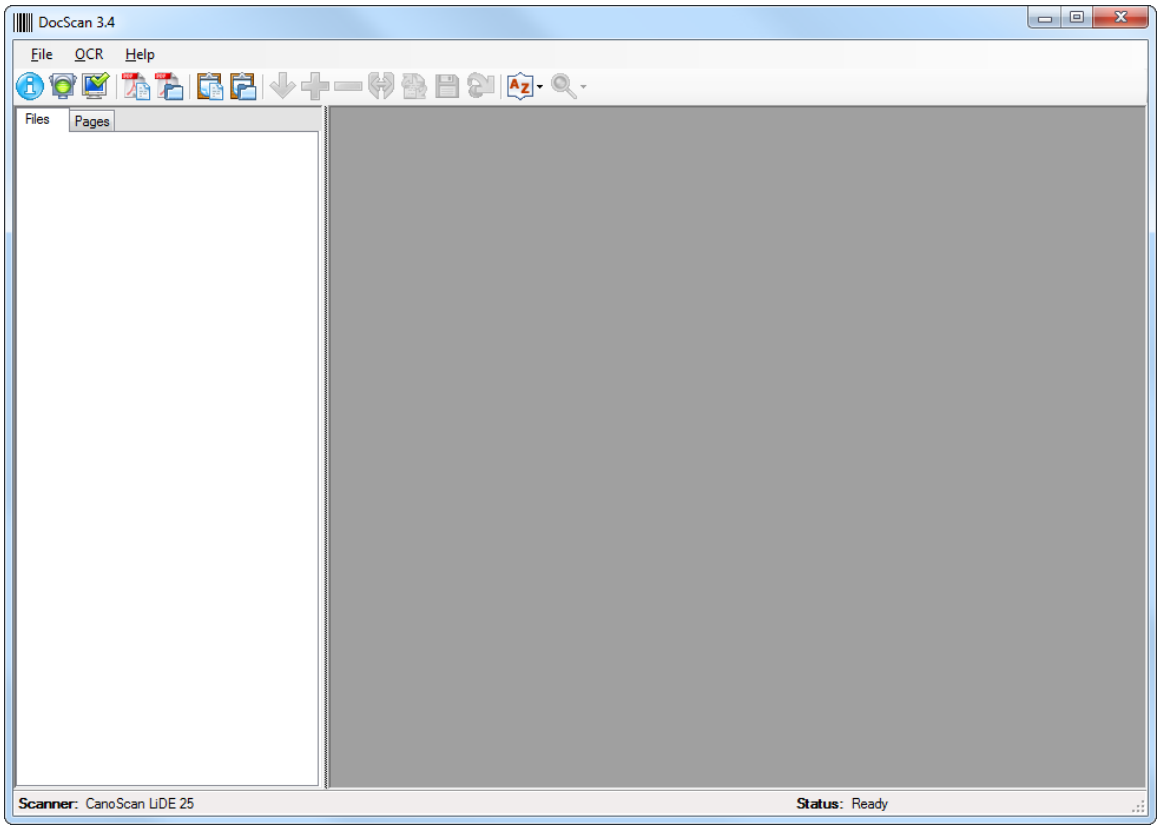


If you fail to select a license file when installing the program, DocScan will show this dialogue screen:



During this trial mode, DocScan will insert a watermark into all scanned images. If this trial period expires (after 45 days), you will only be able to access the "Help" pull-down menu, which provides options for registering and/or activating the software, as necessary. Please consult the guide earlier in this document for information on these processes.

If DocScan is properly registered and activated, the main interface screen will appear as follows:



You will then have to configure the DocScan program according to your scanner and output options.

Please consult the remaining chapters in this guide for information on using and configuring DocScan:

- User Interface
- Scan Process
- OCR, PDF and Forms Processing
- Configuring DocScan

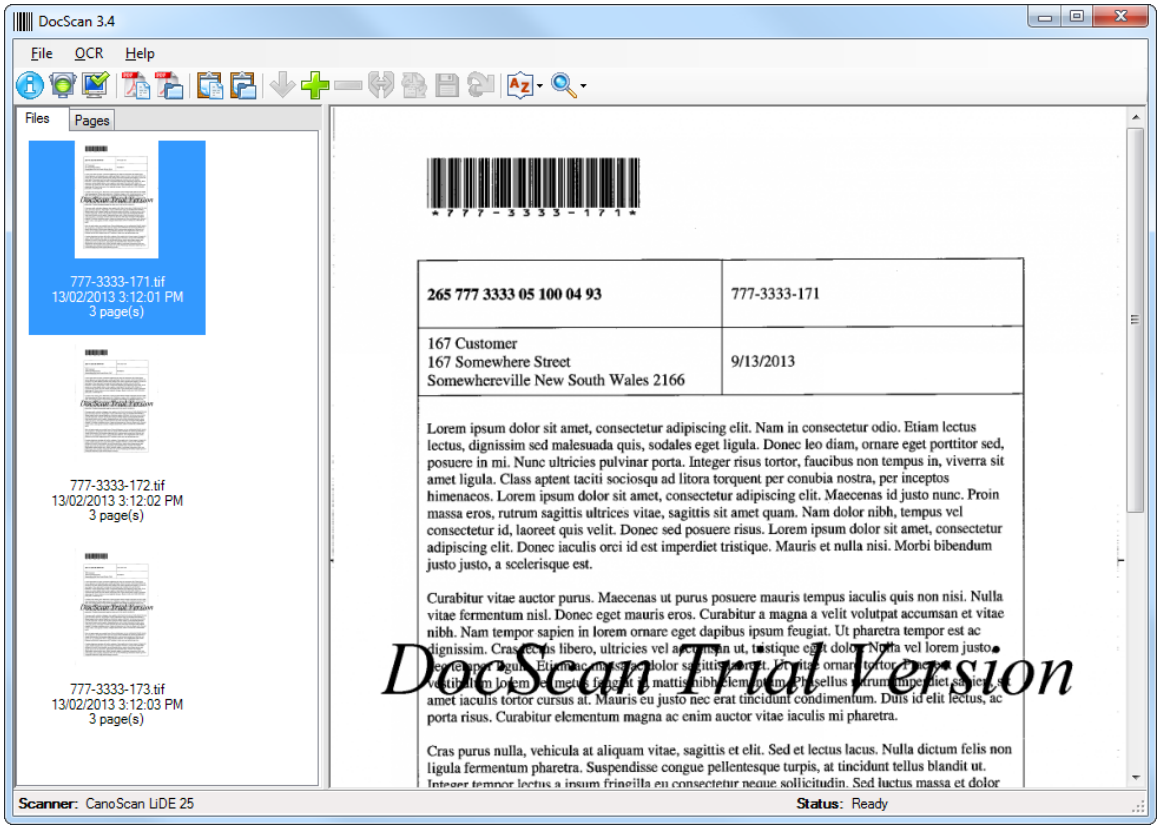
User Manual

User Interface

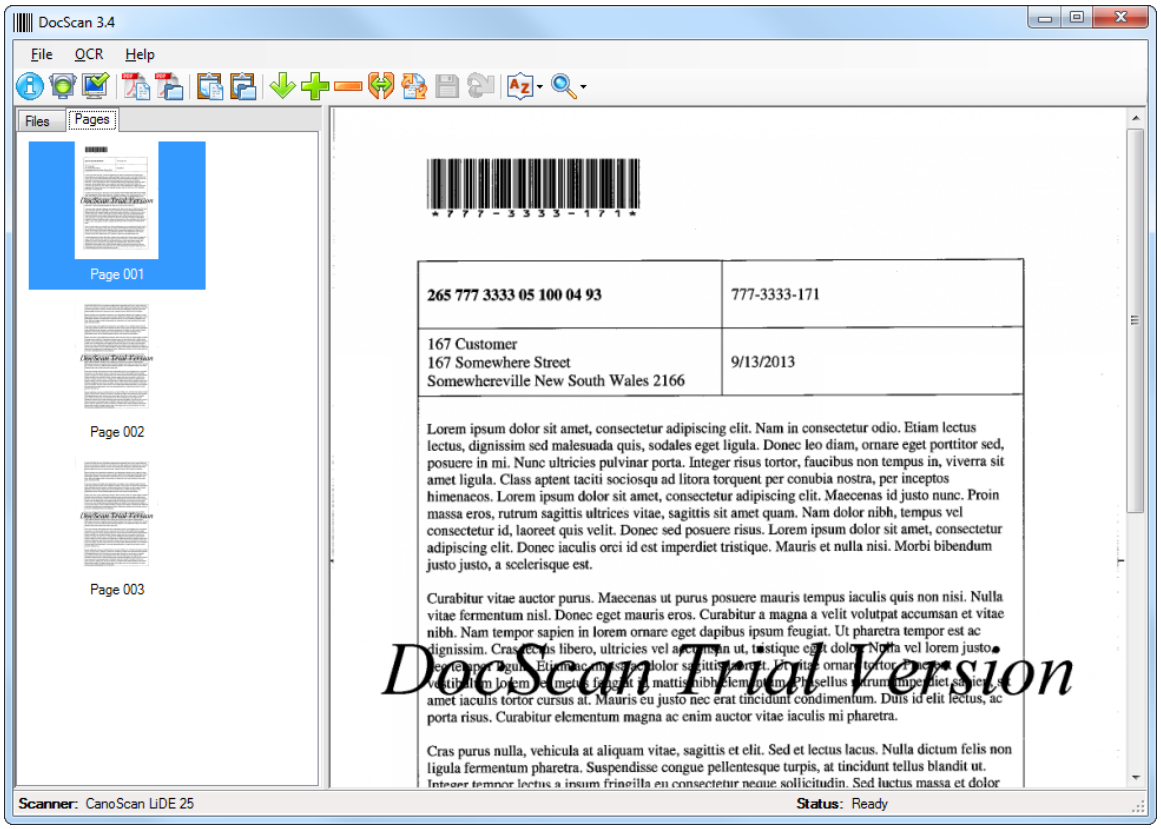
The main DocScan window is divided into two sections — the thumbnail view and the page view. The left section of the window displays a tabbed control containing a File Thumbnail View tab and a Page Thumbnail View tab. The File Thumbnail View tab displays a thumbnail of each file in the Scan Destination folder. (It will display nothing if the Scan Destination folder has not been configured, does not exist or is empty.)

Clicking on a File thumbnail will load the selected file and display the first page in the right section of the screen (page view) for viewing or modifying (see toolbar buttons below). If a File thumbnail is selected, each page of the file will also be displayed as a thumbnail in the Page Thumbnail View tab. Selecting a Page thumbnail will display that page in the page view.

Main window showing folder thumbnail view:



Main window showing page thumbnail view:



Pull-down menus

DocScan has three pull-down menus visible across the top of the page. Some of these functions are duplicated in the toolbar. (Please see the following section for a list of commands available from the toolbar.)

File menu

The File menu contains the following commands:

> *Settings: This option will load DocScan's configuration page. See the Configuring DocScan chapter for more information on the settings that may be changed.*
>
> *View Logs: This option will display the folder location where DocScan saves its log files.*
>
> *DocScan will generate a log file every time a file (or group of files) is processed. These files are of .LOG type, and can be viewed in any plain-text text editor (such as Notepad, for example).*
>
> *DocScan generates one log file per day, with the filename being yyyymmdd.log - so, for example, a log file generated on 12 February, 2013 would have the filename 20130212.log. If more than one scanning session takes place on a single day, DocScan will include all log results in the same file, with each entry being time-stamped.*

Exit: Select this option to exit the DocScan application.

OCR menu
The OCR menu contains the following commands:

OCR the Selected File and Save to PDF: See OCR, PDF and Forms Processing for details.

OCR all Files and Save to PDF: See OCR, PDF and Forms Processing for details.

Forms Process the Selected File and Save to PDF: See Forms Processing section in OCR, PDF and Forms Processing for details.

Forms Process all Files and Save to PDF: See Forms Process section in OCR, PDF and Forms Processing for details.

Create a Forms Processing Template from the Selected File: See Forms Processing Templates for details.

Edit a Form Processing Template using the Selected File: See Forms Processing Templates for details.

Forms Processing Template Designer: See Forms Processing Templates for details.

Help menu
The Help menu allows you to either view information about the specific release of DocScan 3.4 you are running (available by selecting the "About" option), or to view DocScan's online help.

Keyboard shortcuts
There a number of keyboard shortcuts you can use to simplify browsing in DocScan 3.4. In the Files tab, you can scroll up the list of documents by pressing Page Up, and scroll down by pressing Page Down. You can scroll in both directions by using the wheel on your mouse. In the Pages tab, you can scroll up and down the pages of a single document by using the Page Up and Page Down buttons, and the mouse wheel.

Toolbar
Across the top of the main window is the main toolbar, which provides functions for configuration, scanning, editing and viewing.

	**Select TWAIN Source:** This displays a dialogue that allows you to select or change the default scanner. See Configuring DocScan for more details.
	**Start Scan Process:** This commences the scanning and barcode recognition process. See Scan Process for more details.
	**Settings:** This is where various DocScan settings can be configured. See

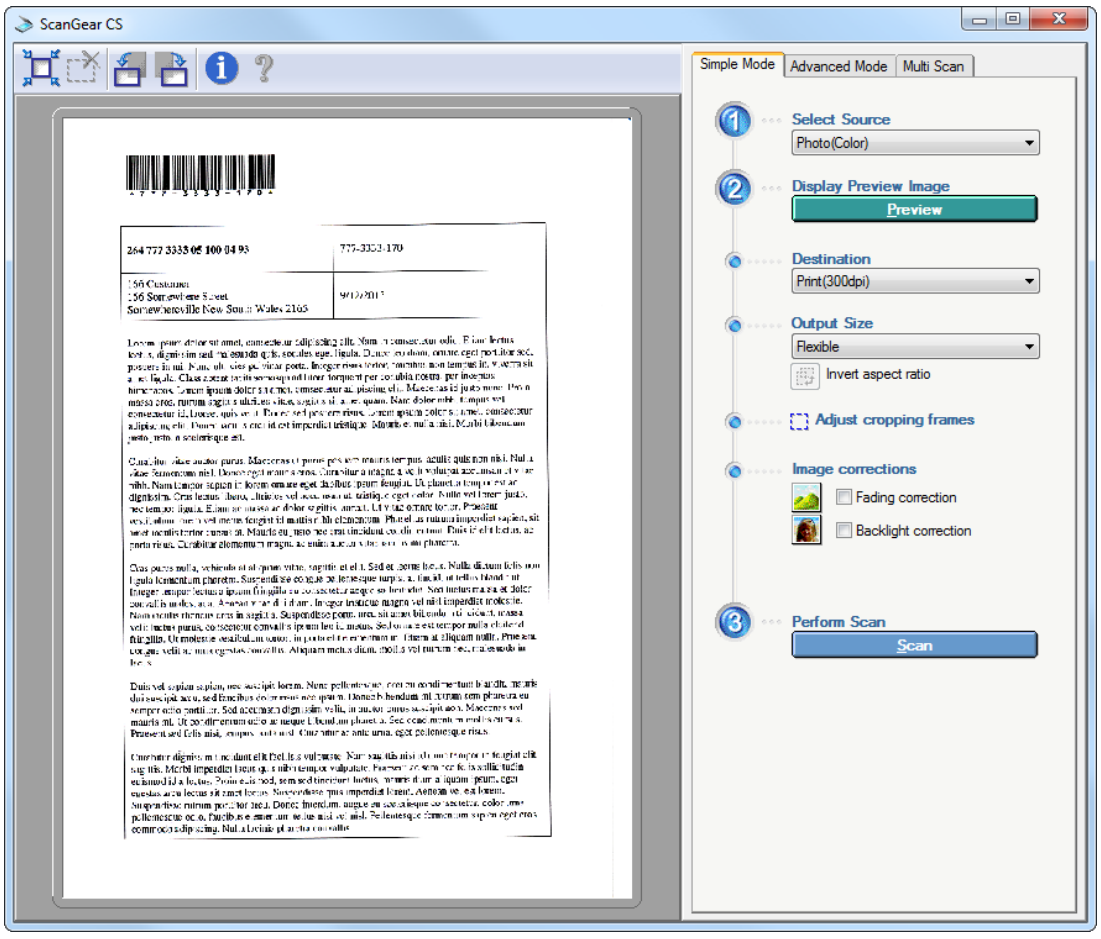
	Configuring DocScan for more details.
	**OCR the Selected File and Save to PDF:** See OCR, PDF and Forms Processing for details.
	**OCR all Files and Save to PDF:** See OCR, PDF and Forms Processing for details.
	**Forms Process the Selected File and Save to PDF:** See Forms Processing section in OCR, PDF and Forms Processing for details.
	**Forms Process all Files and Save to PDF:** See Forms Processing section in OCR, PDF and Forms Processing for details.
	**Insert Page:** This scans and inserts a new page before the current page being viewed in the page viewer. If multiple pages are scanned, multiple pages will be inserted.
	**Append Page:** This scans and appends a new page after the last page of the file being viewed. If multiple pages are scanned, multiple pages will be appended. *Note about Insert and Append Page:* Selecting Insert Page or Append Page will open the scanner's own scan dialogue. This will allow the user to configure options such as resolution, color mode, page size and more. This dialogue will also have a button or menu item to press to commence scanning. When scanning is complete, you may have to manually close this scan dialogue. Since each brand and model of scanner may display a different dialogue, users will need to refer to their scanner's documentation for help in using it.
	Delete Page: This deletes the page currently being viewed in the page viewer.
	**Split:** This function will split the current file into two. You will be prompted for the barcode number of the new file, and the file will be created with the name of *xxxx*.TIF, where *xxxx* is the barcode number. The split will occur on the current page. That is, the new file will consist of the current page and all subsequent pages. All pages that form part of the new file will be deleted from the current file, and both files will be saved automatically.
	**Rotate:** This will rotate the page currently being viewed in the page viewer 180° clockwise.
	**Save File:** This saves changes to the currently selected file.
	**Reload File:** This reloads the current file from disk.
	**Sort Folder View Thumbnails:** This provides a number of sorting options for the file thumbnail view. Available sorting options are: • Date Modified • File Name • Size • Type • Ascending • Descending

Zoom: This provides a number of zooming options for the page view.

Scan Process

Press the Start Scan Process button to begin the scanning and barcode recognition process. The scan process follows these steps:

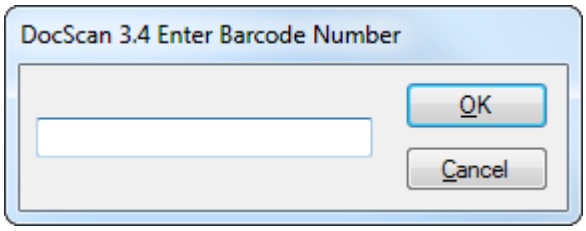
1. The scanner's own "scan" dialogue will open. This will allow you to configure options like resolution, colour mode, page size and more. This dialogue will also have a button or menu item to press to commence scanning. Here is the scan dialogue for a Canon flatbed scanner:



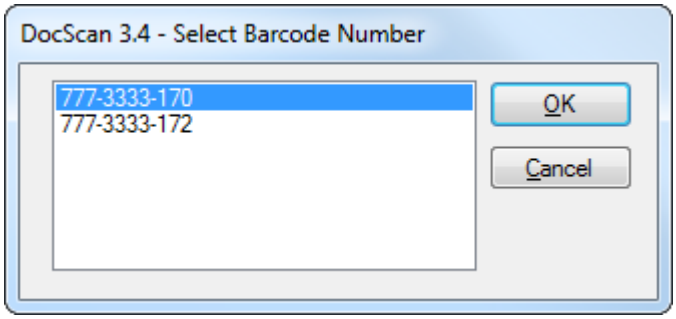
Since each brand and model of scanner will display a different dialogue, you will need to refer to your scanner's documentation for help in using it.

2. The first page is scanned.

3. If the page contains a barcode, a new file will be created in the Scan Destination directory with the filename *xxxx*.TIF, where *xxxx* is the barcode number. If no barcode is found, you will be asked if you wish to use the saved barcode number (if one exists), or you will be prompted to manually enter a barcode number for the new file as follows:



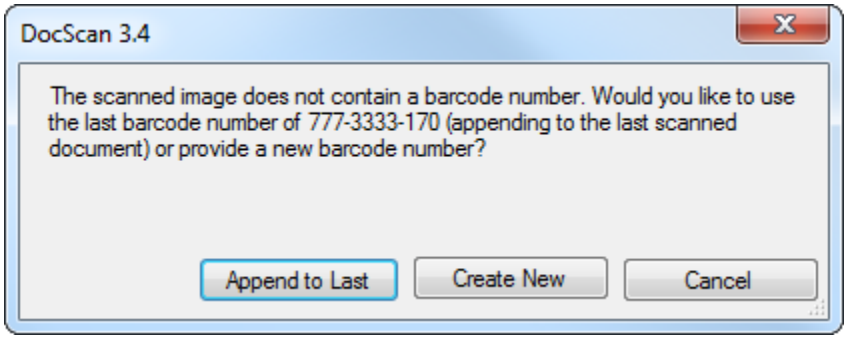
If multiple barcodes are found, you will be shown a list of the recognized barcode numbers and be asked to select one to use for the new file. For example:



The scanned page will become the first page of the new file and the barcode number will be saved to be used for all subsequent images without a barcode number. (This continues until another barcode number is found.

4. The next page is scanned.

5. If the page does not contain a barcode, it is appended to the current file. If there is no current file, the saved barcode number from Step 1 is used.



6. If the page does contain a barcode, a new file is created, as in Step 1.

7. Steps 3 to 5 are repeated until all pages are scanned, the scan process is canceled, or an error occurs.

8. When scanning is complete, the scanner's dialogue window may need to be manually closed.

OCR, PDF and Forms Processing

Beginning with version 3.0, DocScan has the ability to perform Optical Character Recognition (OCR) on a scanned document. This feature will allow you to process a document once it has been scanned, after which DocScan will convert the raw image data of the scanned document into machine-readable text. With this extracted text, you would then have the ability to index documents within RecFind or "copy and paste" the text from the image into a word processor, for example.

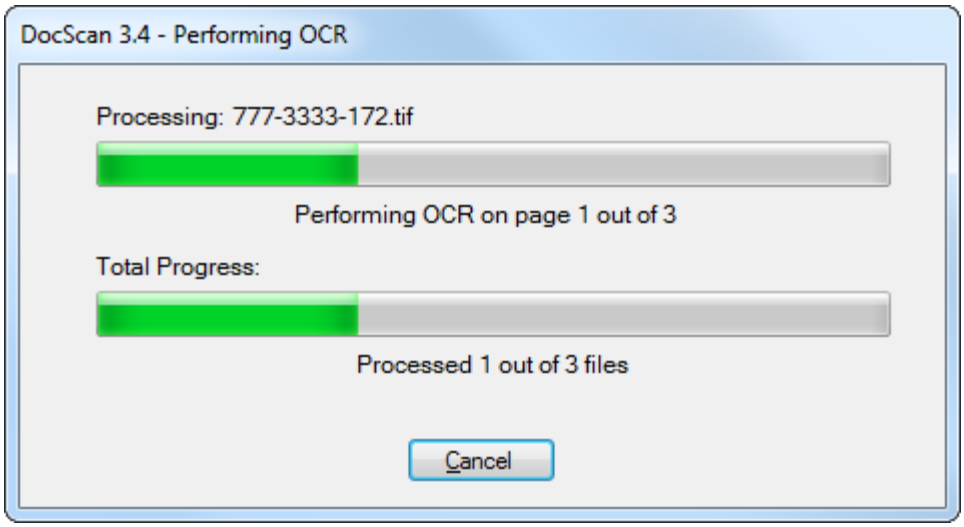
To perform OCR on a scanned document, you will need to click either the "OCR all Files and Save to PDF" or "OCR the Selected File and Save to PDF" option in the OCR pull-down menu, or the appropriate icons on the main toolbar. (See the User Interface chapter for more information on this.)

The process of performing OCR on a document and then converting it to PDF is "atomic" — that is, it is not possible to separate one from the other.

There are two ways in which OCR and PDF conversion can be performed:

OCR Selected File and Save to PDF
This option is achieved by pressing the "OCR Selected File and Save to PDF" button on the main DocScan window. To use this option, you must first have selected a document in the "Files" thumbnail view. (For more information on the DocScan main window, see the User Interface chapter.)

Once you have selected a file and pressed the button, DocScan will initiate the OCR process, you will see a dialogue box similar to that shown below:



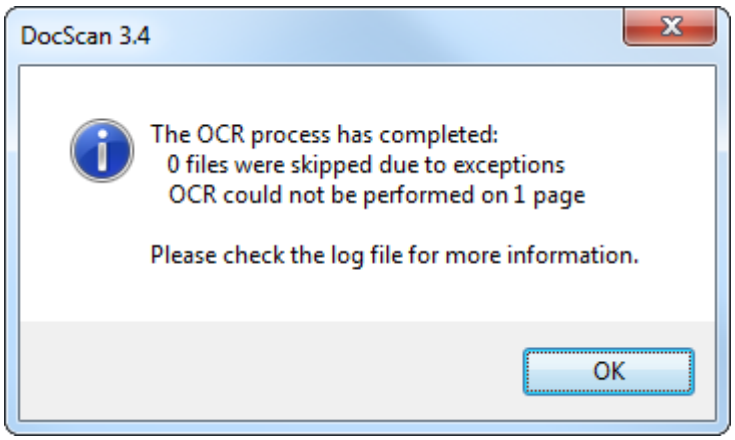
Once the conversion process is complete, DocScan will save the converted PDF to the same location as the TIFF file. The file will have the same name as the original TIFF image, with the extension ".pdf". In the example shown above, the resultant file will be called "777-3333-172.pdf".

Note: it is possible to configure DocScan so that it will save the resultant PDF file to a different location. Please see the Configuring DocScan chapter for information on how to change this setting.

OCR all Files and Save to PDF
This option is similar to that described above, although it will convert every image in your nominated directory. As above, all converted PDF files will appear in the same directory, and will share the same name as their respective source files; only the extension will change.

Note: there are a number of settings that determine how DocScan will process and treat files during the OCR/PDF conversion process. Please see the Configuring DocScan chapter (and the "OCR + PDF Options" section) for information on these settings.

If there were errors, the following dialog box will be shown when the operation is complete:



This provides a summary of the conversion process. Any errors are described in detail in a DocScan log file. For information on accessing log files within DocScan, please see the "Pull-down menus" section of the User Interface chapter.

OCR and image resolution
We recommend that you provide source images at 300 dots per inch (dpi) resolution. You may provide images at resolutions higher than this value, although it is possible that excessively high resolutions will cause the DocScan OCR engine to exhaust the available memory. If this happens, an error will be generated and noted in the log file.

This error appears in the log file as follows:

> ERROR: error [*x*]
> Could not perform OCR on page *x*. Please check the user manual for more information.

There are a number of reasons that this error may occur.

The source TIFF image:

- may be too large
- may be too small
- may be corrupt
- may not contain any text
- may have been created using a non-standard compression method (e.g. Deflate)

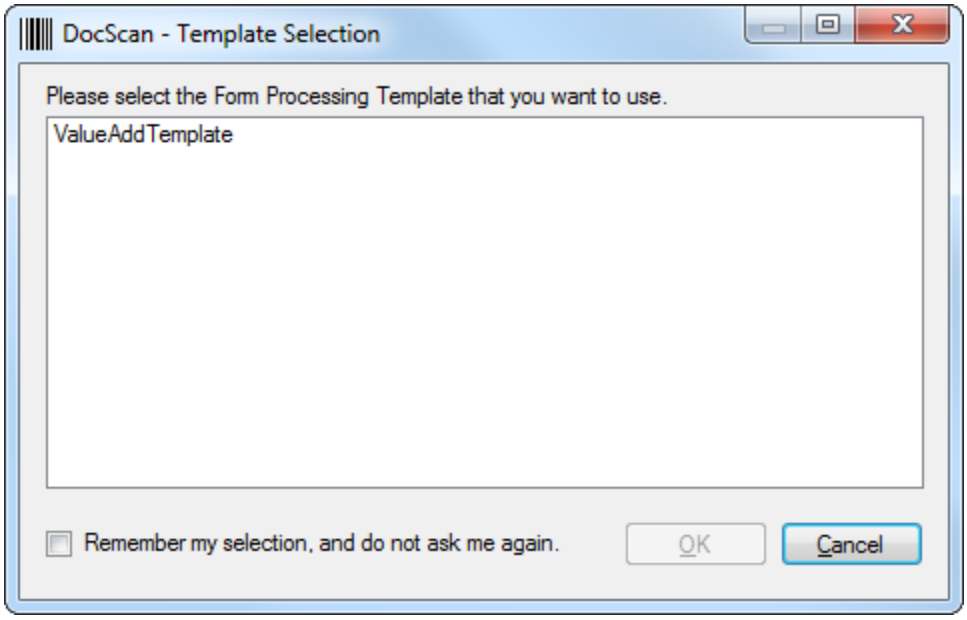
In these cases, we recommend that you re-scan the document at a more accommodating resolution. (For information on modifying the settings of your scanned documents, please refer to the information supplied by the manufacturer of your scanning hardware.)

If an error occurs on a single page (or collection of pages) within a multi-page document, that page (or those pages) will be inserted into the PDF file as images, but there will be no embedded text behind them. If a single-page document fails, the resulting PDF file will contain the image, but no embedded text.

Forms Processing
To perform Forms Processing on a scanned document, you will need to click either the "Forms Process all Files and Save to PDF" or "Forms Process the Selected File and Save to PDF" option in the OCR pull-down menu, or the appropriate icons on the main toolbar. (See the User Interface chapter for more information on this.)

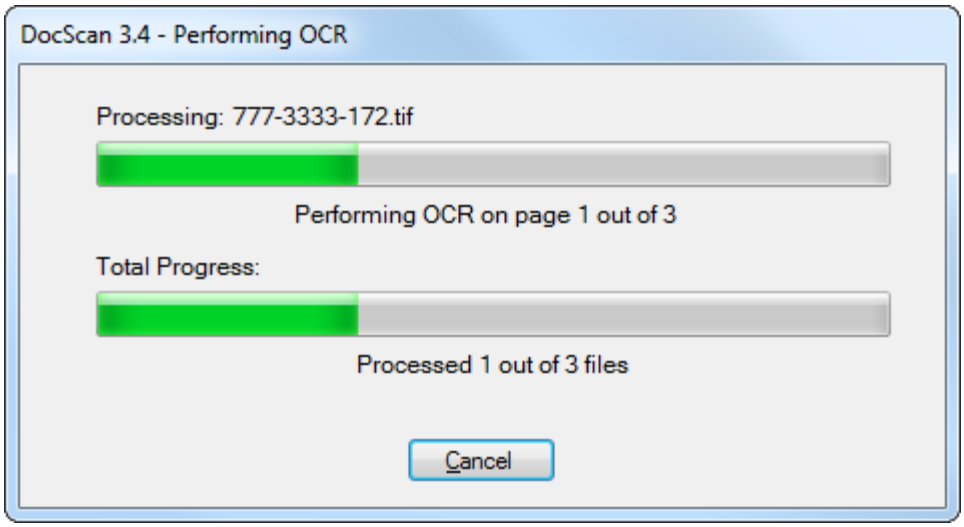
Forms Process Selected File and Save to PDF
This option is achieved by pressing the "Forms Process Selected File and Save to PDF" button on the main DocScan window. As with "OCR the Selected File and Save to PDF" you must first have selected a document in the "Files" thumbnail view.

Once you have selected a file and pressed the button, DocScan will initiate the OCR process. If you have configured DocScan to "Ask Me Which Template To Use" in the settings (see Configuring DocScan for more detail), then you will be prompted to select a template to use.



Select a template by either clicking on a template and pressing the OK button, or by double-clicking on a template in the list. If you have no templates, you should create one first using the template designer (see the Forms Processing Templates section of this manual). If you want to always use a particular template, select it, and then tick the "Remember my selection" tick box. This will configure DocScan to use the selected item as the default template (see Configuring DocScan).

If you have chosen a template then the process will continue and you will see a dialogue box similar to the one shown below:



Once the conversion process is complete, you will notice additional files have been created along with the PDF file. The number and structure of the XML files will depend on your chosen method of import into the RecFind database, either RecScan or Xchange.

Forms Process all Files and Save to PDF
This option is similar to "Forms Process the Selected File and Save to PDF", except it will process every image in your nominated directory.

RecScan
The default method of import on first use of DocScan is RecScan. For this method there will be an information file for each PDF, the additional file will have the same name as the PDF file except it will have an XML extension. This XML file will have the all the information regarding the OCRed text that was contained in each ROI (region of interest).

Example of the information file created for RecScan:

```
<?xml version="1.0" encoding="utf-8" ?>
- <Document DocumentName="PersonList.xml">
  - <Page PageNumber="1">
    - <Region RegionName="DATE">
        <Text>07/09/2009</Text>
      </Region>
    - <Region RegionName="PARTNUMBER">
        <Text>9420</Text>
      </Region>
    - <Region RegionName="EMAIL">
        <Text>S.DIAS@EXERCISES.COM</Text>
      </Region>
    - <Region RegionName="EXTERNALID">
        <Text>AMANDA BLAIR</Text>
      </Region>
    </Page>
  - <Page PageNumber="2">
    - <Region RegionName="ENTITY">
        <Text>Entity</Text>
      </Region>
    </Page>
  - <Page PageNumber="4">
    - <Region RegionName="MELISSA">
        <Text>TURNER, MELISSA</Text>
      </Region>
    </Page>
  </Document>
```

You will notice that the information is separated by page number and then by the regions contained on each page of the template used.

RecScan will use the information XML file to match region names to fields in the EDOC table. For example, any text found in the region named "Abstract" will automatically be entered in the Abstract field of an EDOC record when PDF file is uploaded to the RecFind database.

Xchange

If your chosen method of import is Xchange then only a single XML file is created. The name of this file can be specified in the Form Processing Options under the Xchange Options. This single file will have all the information regarding the OCRed text that was contained in each ROI in addition to an encoded copy of the PDF file. Xchange will only use the XML file for import and does not need the PDF file that is also created.

Example of the information file created for Xchange:

```
<?xml version="1.0" encoding="utf-8" ?>
- <ExportSet table="Import" xmlns:xs="http://www.w3.org/2001/XMLSchema">
  - <xs:schema id="ExportSet" xmlns:msdata="urn:schemas-microsoft-com:xml-msdata">
    - <xs:element name="ExportSet" msdata:IsDataSet="true" msdata:Locale="en-AU">
      - <xs:complexType>
        - <xs:choice maxOccurs="unbounded">
          - <xs:element name="Record">
            - <xs:complexType>
              - <xs:sequence>
                  <xs:element name="EXTERNALID" type="xs:string" minOccurs="0" xmlns:xs="http://www.w3.org/2001/XMLSchema" />
                  <xs:element name="REGISTRATIONNUMBER" type="xs:string" minOccurs="0" xmlns:xs="http://www.w3.org/2001/XMLSchema" />
                  <xs:element name="DUEDATE" type="xs:string" minOccurs="0" xmlns:xs="http://www.w3.org/2001/XMLSchema" />
                  <xs:element name="ADDRESSEE" type="xs:string" minOccurs="0" xmlns:xs="http://www.w3.org/2001/XMLSchema" />
                  <xs:element name="EDOCEXTERNALID" type="xs:string" minOccurs="0" xmlns:xs="http://www.w3.org/2001/XMLSchema" />
                  <xs:element name="IMAGE" type="xs:base64Binary" minOccurs="0" xmlns:xs="http://www.w3.org/2001/XMLSchema" />
                  <xs:element name="FILENAME" type="xs:string" minOccurs="0" xmlns:xs="http://www.w3.org/2001/XMLSchema" />
                  <xs:element name="SUFFIX" type="xs:string" minOccurs="0" xmlns:xs="http://www.w3.org/2001/XMLSchema" />
                  <xs:element name="PUBLISHEDDATE" type="xs:dateTime" minOccurs="0" xmlns:xs="http://www.w3.org/2001/XMLSchema" />
                  <xs:element name="ORIGINALPATH" type="xs:string" minOccurs="0" xmlns:xs="http://www.w3.org/2001/XMLSchema" />
                </xs:sequence>
              </xs:complexType>
            </xs:element>
          </xs:choice>
        </xs:complexType>
      </xs:element>
    </xs:schema>
  - <Record>
      <EXTERNALID>264777333305 1000493</EXTERNALID>
      <REGISTRATIONNUMBER>7773333-172</REGISTRATIONNUMBER>
      <DUEDATE>09/03/2013</DUEDATE>
      <ADDRESSEE>Knowledgeone Corporation Level 5 56 Berry Street North Sydney 2060</ADDRESSEE>
      <EDOCEXTERNALID>777-3333-172.pdf</EDOCEXTERNALID>
      <IMAGE>JVBERi0xLjQKJeLjz9MKMSAwIG9iago8PC9GaWx0ZXIvRmxhdGVEZWNvZGUvTiAzL0xlbmd0aCAyNTk2Pj5zdHJlYW0KeJydlndUU9kWh8+
      <FILENAME>777-3333-172.pdf</FILENAME>
      <SUFFIX>pdf</SUFFIX>
      <PUBLISHEDDATE>2013-02-08T16:37:17.525+11:00</PUBLISHEDDATE>
      <ORIGINALPATH>[SCAN01] C:\Scanned Documents</ORIGINALPATH>
    </Record>
  </ExportSet>
```


You can see that the file has a schema that defines the elements to be used in the import by Xchange. The actual information from the region and the encoded image of the PDF is included under the *Record* tag.

This XML file can be used as a data source for an import in Xchange into the Recfind database. Please see the relevant documentation for more information about using Xchange to import into the Recfind database.

Forms Processing Templates

DocScan 3.4 has the ability to **read** a scanned document and extract the relevant data to a computer-readable format. To use the forms processing functionality you must first create a forms processing template that defines a document's regions of interest (ROIs). To do this you will use the Forms Processing Template Designer. The Designer is accessed through the OCR menu on the main screen of DocScan.

There are three options when you are opening the Template Designer:

- *Create a Forms Processing Template from the Selected File*
- *Edit a Forms Processing Template using the Selected File*
- *Forms Processing Template Designer*

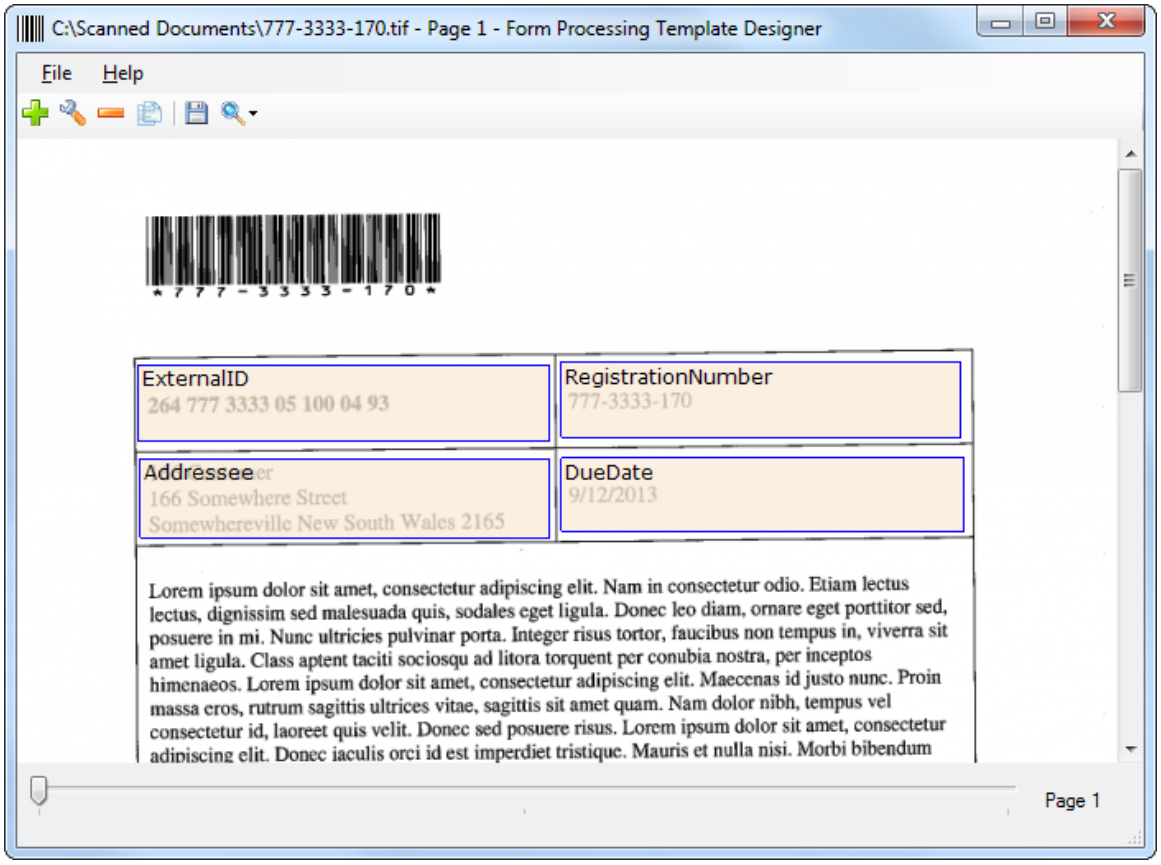
Create a Forms Processing Template from the Selected File
Use this option to create a new template for an existing tiff image. Select an image from the thumbnail view then choose this option. The Template Designer will be opened with the Tiff

you have selected. You can then create the required ROI and save your template. See the Toolbar section for more details on adding ROIs.

Edit a Forms Processing Template using the Selected File
If you wish to make changes to an existing template this use this option. Again, select an image from the thumbnail view making sure that you chose one that matches the template you wish to edit. The designer will open with the selected Tiff, you will then be prompted to select the template you wish to edit. Once you have selected a template the regions of interest will be loaded and you can make any required changes.

Forms Processing Template Designer
Select this option to open a blank template designer. If you have a blank designer, you will first need to load a tiff image into the designer. Do this by selecting **Load TIFF Image** from the File menu. You can then create the required ROI and save your template. See the Toolbar section for more details on adding ROIs.

Once you have made a selection you will then be shown a variant of the following screen:



Depending on the option you have selected a tiff and ROIs may or may not be shown.

Pull-down menus

The DocScan Forms Processing Template Designer has two pull-down menus visible across the top of the page. Some of these functions are duplicated in the toolbar. Please see the Toolbar section below for a list of the commands that are available.

File menu
The File menu contains the following commands:

Open Template: This option will allow you to load an existing template into the Template Designer.

Note: Any Regions that are currently shown in the designer will be cleared when the template is loaded.

Load Tiff Image: This option will allow you to load a Tiff image into the Template Designer.

Note: A Tiff must be loaded before any Regions can be added or existing templates loaded.

Merge Template: Select this option to merge the Regions from a saved template to the ones that are already in the designer.

Close: Select this option to close the Template Designer.

Save: Select this option to save any changes to the currently open template.

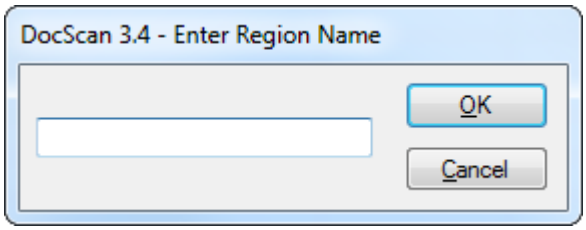
Save As: Select this option to save any changes to the currently open template under a new name.

Help menu
The Help menu allows you to either view information about the specific release of DocScan 3.4 you are running (available by selecting the "About" option), or to view the contents of DocScan's help.

Toolbar
Across the top of the designer window is the toolbar, which provides functions for template creation and modification.

Add: To add an ROI, you will need to click the this icon, and then drag your mouse to cover an area of the document.

After selecting an area of the document, you will be asked to specify a name for the region:



For more information on naming regions see Region Names

Please note that it is better to make a region slightly larger than what may apparently be required, so as to account for differences between each scanned document. So long as the first letter of a word falls within a region, the entire word (that is, all letters until a "space" character is encountered) will be captured.

Modify: To rename a region, you can select the region and click this icon. Alternatively, double-click the region and enter the new name in the text box that appears.

Delete: To delete a region, select the region with the mouse and click this icon. Alternatively, you may highlight region and then press the Delete key. To select multiple regions prior to delete, hold down Ctrl while highlighting regions. This will allow you to delete more than one region at a time.

Clone: To clone a region, select the region to copy and then click this icon. This will create the region on all pages of the document. Each region will be given a unique name by appending the page number to the name of the region being cloned.

Save: To save the template, press the this icon, or select Save or Save As from the File menu. If it is a new template, you will be prompted to save the template to disk. It is recommended to save the template to the default folder or the folder you have designated as the "template folder" in the DocScan configuration.

See Configuring DocScan for more information.

Zoom: To change the way the TIFF image is displayed in template designer, use this icon and select the preferred ratio.

Region Names

Please consider the following when naming regions in the template designer:

- Region names cannot contain any whitespace characters.
- Each region name must be unique within a given template.

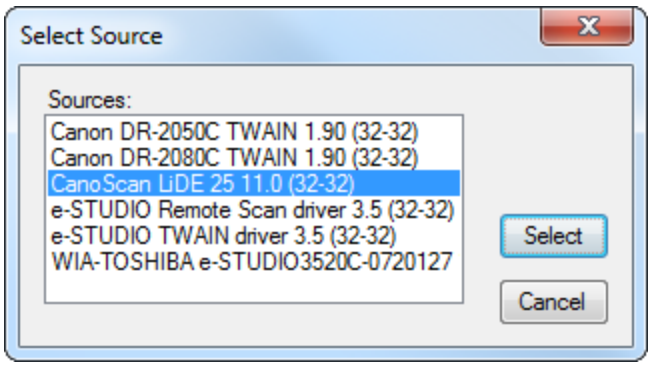
If your chosen method of import into RecFind 6 is to use RecScan then these region names are used by RecScan to link to corresponding EDOC fields. For example, any text found in the region named "Abstract" will automatically be entered in the Abstract field of an EDOC record with the TIFF image or PDF document is uploaded to the database.

For information on how forms processing works when converting to PDF, see the OCR, PDF and Forms Processing section of this manual.

Configuring DocScan

Before you can start using the scanning features of DocScan, you will need to configure the following settings.

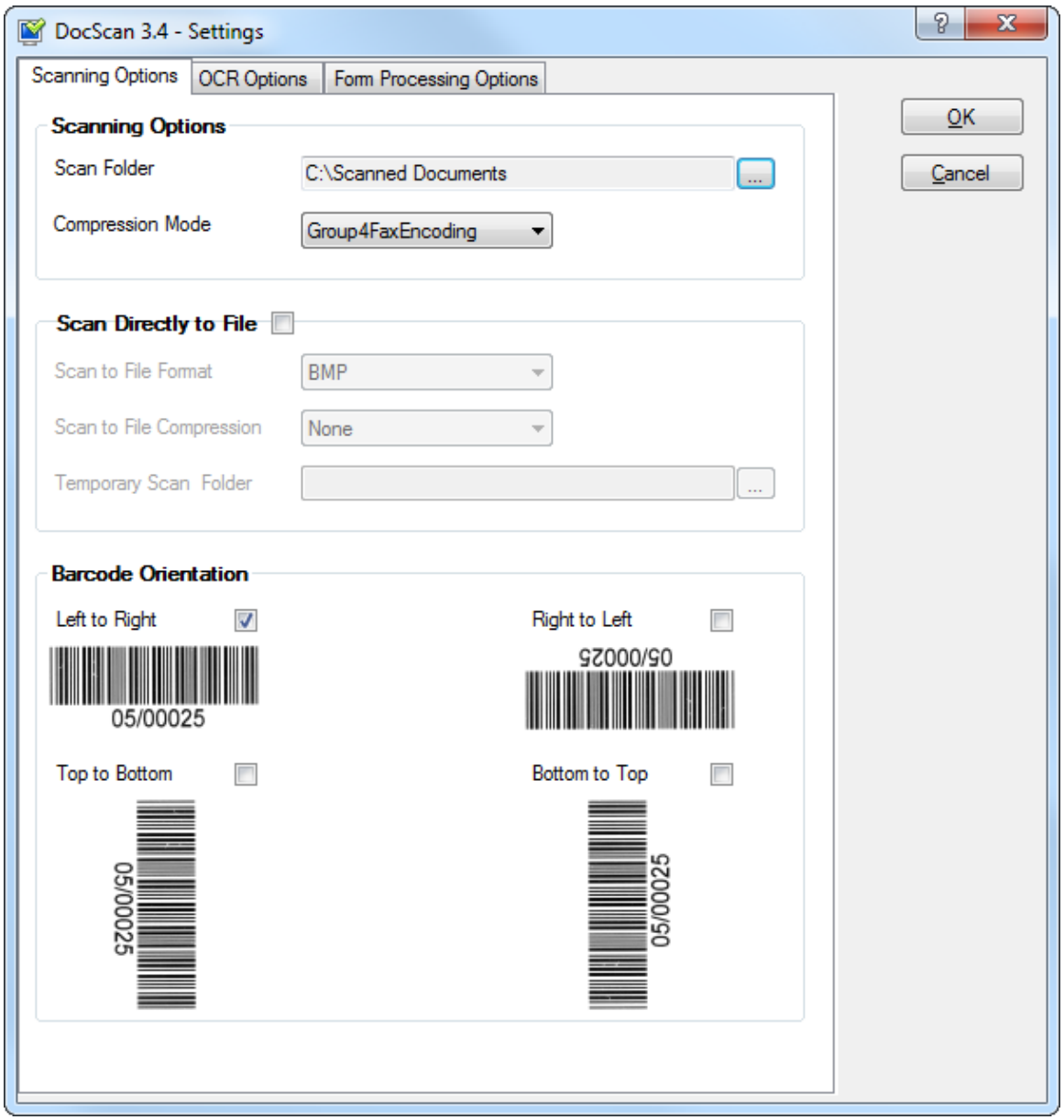
Select your scanner
Press the Select TWAIN Source button to display a list of the scanners that are installed on your computer. A dialogue box similar to the following will open:



Note that the contents of the Sources list will depend on the scanners you have installed on your PC, and may differ from the list shown above. If there are no scanners shown, you will need to ensure that your scanner's software is properly installed and that the scanner is connected and turned on.

Select the scanner that you wish to use and press the Select button. This will save your selection as the default scanner for future use.

Configuring Scanning Options
Press the Settings button to display the Settings window shown here:



Scan Folder: This is where DocScan will save scanned images as multi-page TIFF files.

Compression Mode: Scanned image files can be compressed to save disk space. Although the default of Group4FaxEncoding will be suitable for most situations, you can choose from any of the following Compression modes:

- no compression
- Group 3 Fax Encoding
- Group 4 Fax Encoding
- JPEG compression
- Macintosh Packbits
- Deflate

- LZW
- Modified Huffman

In addition to the two settings described above, DocScan provides a number of other settings. These settings will not need to be configured for most situations, but you may need to use them if you have memory usage or barcode recognition problems.

Scan Directly to File: Some scanners support scanning directly to file, which may reduce memory usage. If this option is supported by your scanner, and it is selected, images will be scanned to a temporary file instead of to the computer's memory. Selecting this option may help if you experience performance issues or errors due to excessive memory use. If this option is not supported by your scanner, or if this does not resolve your problems, you will need to reduce the quality of the scanned images.

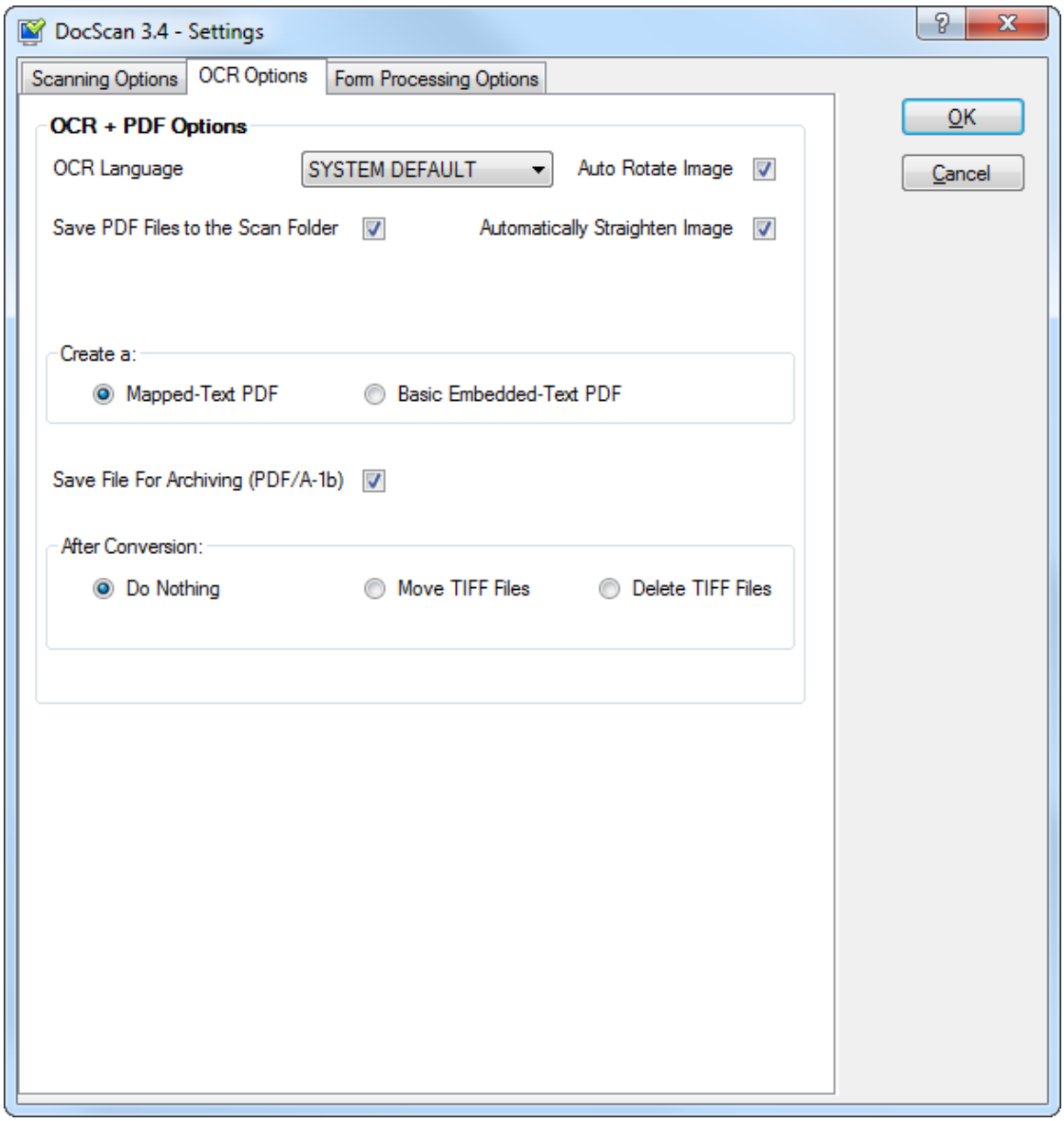
Scan to File Format: This option selects the format that the temporary file will be saved in. DocScan provides the formats listed below, but please be aware that not all scanners support all formats. DocScan will notify you if the chosen format is not supported by your selected scanner. The available formats are:

- TIFF
- PICT
- BMP
- XBM
- JFIF
- FPX
- TIFFMULTI
- PNG
- SPIFF
- EXIF

Temporary Scan Folder: This is the folder where the Scan Directly to File setting will save the temporary files during scanning. DocScan will delete the files once they are no longer required.

Barcode Orientation: DocScan can read barcodes both horizontally and vertically, in addition to back-to-front barcodes, but only the most common option of horizontal, left to right, is selected by default. If your barcodes are placed differently, please select the appropriate option. You can select more than one option, but please note that there may be an accompanying drop in performance.

Configuring OCR and PDF Options



Note: this option will only be visible if you have Microsoft Office Document Imaging (MODI) installed on your system. For more information on MODI and its role in DocScan, please see the "Requirements" section of the Welcome chapter.

OCR Language: This option allows you to select the language that will be employed in determining the characters read in by the OCR process. There are a number of languages that can be selected, as well as the option "SYSTEM DEFAULT". If you select this option, DocScan will use the language setting that you currently employ in your Microsoft Windows installation.

Auto-Rotate: This option will automatically rotate the image so that the orientation is correct. This is important for any optical character recognition that is to be performed.

If an image is rotated ninety degrees from "upright" (you may be scanning in documents in landscape orientation, for example), the OCR parser will not be able to determine letter values. By automatically rotating the image to the upright position, you will be able to ensure that OCR is performed on all pages correctly. Once the image is automatically rotated, the source TIFF image will be updated to reflect these changes. Please note that the auto-rotate option may have less than satisfactory results if the source images are scanned at a resolution of less than 300 DPI.

Save PDF files to the Scan Folder: This checkbox allows you to determine the location of the converted PDF files. If the checkbox is selected (which it is by default), the PDF files will be saved to the same location as the scanned TIFF images. If the checkbox is not selected, you will be asked to specify an alternative directory. You may then browse to a folder or network location that is more suitable to your needs.

Automatically Straighten Image: If this option is selected, DocScan will check to see if any scanned images are a few degrees "off-centre". If this is the case, DocScan will correct the skew when performing the OCR operation, in an attempt to produce a more accurate assessment of the characters contained in the document. Once DocScan has made the modification to straighten the image, this save will be changed in the source TIFF file (and the straightening will also be preserved in the PDF file that is generated); that is, DocScan will change the source file.

Note: If the source image is scanned at a resolution of greater than 900 dots per inch, the output PDF file's resolution will be set at a maximum of 900 dots per inch if this option is selected.

Create a Mapped Text PDF/Basic Embedded-Text PDF: This option allows you to select the type of PDF generated by DocScan. The two options are as follows:

> *Mapped-Text PDF: this option will insert scanned text at the correct locations in the PDF file (that is, it will preserve formatting styles present in the original document). This option makes it simple to copy-and-paste selections of text from the resultant PDF file. It is the slower of the two options, as it requires more computational power. This setting is required to use the Forms Processing feature of DocScan.*
>
> *Basic Embedded-Text PDF: this option simply inserts the extracted text in paragraphs, starting at the top-left corner of the document. This option requires significantly less computation time, and would benefit users who will index/search and copy all text from a page (and not just a selection of it).*
>
> "Basic Embedded-Text PDF" is selected by default, as it produces files that are smaller than a mapped-text document. Switch this setting to "Mapped-Text PDF" if document size is less of a consideration, and you want to more accurately replicate the formatting of the original document. However, the Forms Processing feature of DocScan will not work with the Basic Embedded-Text PDF feature turned on.

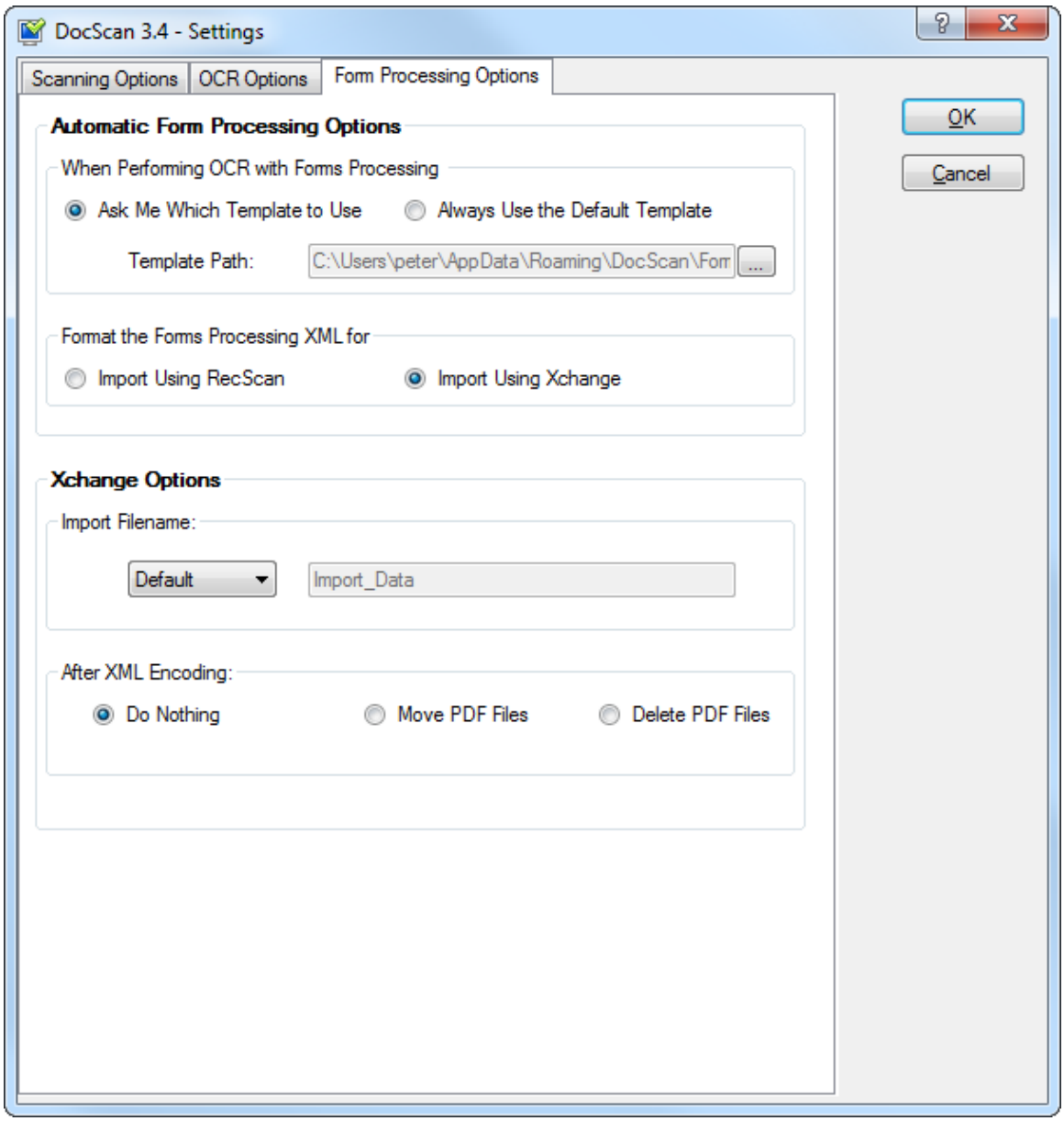
Save File For Archiving (PDF/A-1b): This checkbox allows you to specify that all PDFs created by DocScan should meet the PDF/A-1 standard. If selected, all PDFs will be created with Level B conformance (PDF/A-1b) to the standard. For scanned documents, conformance with PDF/A-1b is sufficient, even if they are processed using OCR to created a text-searchable PDF.

> *Note:* PDF/A-1b (Level B Conformance) meets the minimum requirements of the PDF/A-1 standard.

After Conversion: This option allows you to determine what will happen to the source TIFF images once the conversion is complete.

> *Do Nothing:* this will leave the TIFF images where they are, in the source folder.
>
> *Move TIFF Files:* if you select this option, you will be presented with the option to select a folder path. Once you do this, DocScan will move the images to this folder once the conversion process is complete.
>
> *Delete TIFF Files:* if this option is selected, DocScan will delete the images from the file system once the conversion process is complete.

Configuring Form Processing Options



Note: These options will only be enabled when the Mapped-Text PDF option is selected under OCR Options

When Performing OCR with Forms Processing: This option allows you to select the method of selecting a template to be used when forms processing occurs.

Ask Me Which Template To Use: with this option selected (and Forms Processing enabled on main form), you will be prompted to select a template to use each time you convert scanned images to PDF. You can select a directory in the "template directory" text box where your list of available templates will load from.

Always Use The Default Template: if you select this option, then the selected "Default template" will always be used when converting scanned images to PDF.

Format the Forms Processing XML for: The information that is contained in the regions of interest and retrieved as part of the OCR process will be presented in an XML file. This option allows you to select the format of the XML depending on how you wish to import the information into the RecFind database.

Import Using RecScan: with this option selected an additional file will be created with each PDF file, it will have the same name as the PDF file but will have an XML extension. The XML will be formatted for import with RecScan into the EDOC table in the RecFind database.

Import Using Xchange: if this option is selected, a single XML file is created, the name of which can be set in Xchange options. The file will contain all information retrieved during OCR as well as an encoded version of the PDF. This XML file can be used as a data source for Xchange to import the information to the RecFind database. This method gives you the ability to add to any table (including custom tables) in the database.

Xchange Options
If you choose to use Xchange as your method of import then the following additional options will need to be set.

Import Filename: This section allows you to specify the filename for the XML file that will be created by DocScan for import into Xchange. There are three options:

Default: The filename will be *Import_Data.xml*

Timestamp: The filename will be the same as the default option but will have a timestamp added, *Import_DatayyyyMMddHHmmss.xml*

Custom: with this option you will be able to enter any filename of your choice into the available textbox.

After XML Encoding: This option allows you to determine what will happen to the PDF files once forms processing is complete.

Do Nothing: this will leave the PDF files where they are, in the source folder.

Move PDF Files: if you select this option, you will be presented with the option to select a folder path. Once you do this, DocScan will move the files to this folder once forms processing is complete.

Delete PDF Files: if this option is selected, DocScan will delete the files from the file system once forms processing is complete.