TIFF Junction
Reference Guide Version 3.2

Standard and Professional Editions

Version 3.21 August 2010

© Copyright 2001 - 2010 Aquaforest Limited

http://www.aquaforest.com/

CONTENTS

1 INTRODUCTION

TIFF Junction provides a set of processing tools for TIFF files, including splitting, merging, conversion to PDF and setting of security and other attributes of the converted PDF files. The Professional Edition additionally provides the ability to create Searchable PDFs from TIFF files and Image-Only PDF files, and to generate OCR text files.

The product has both a GUI and command-line interface. Batch processes can be defined using the GUI to create an XML Job Ticket file which can be run using the GUI, or using the command line.

1.1 System Requirements

- Windows 2003, 2008, Windows XP, Windows Vista, Windows 7
- Approximately 140Mb of disk space is required.
- To use the WSH Scripts WSH 5.6 or later is recommended. This will be installed automatically by the install program if it is not present. It can also be downloaded from http://msdn.microsoft.com/scripting if necessary for any reason.

1.2 Supported TIFF File Formats

- CCITT Group 3 (1-D), Group 3 (2-D)
- CCITT Group 4
- CCITT RLE
- JPEG (Type 6 Compression)
- Uncompressed (Bitonal)

1.3 Installation

The setup.exe installation script will guide you through the process of setting up TIFF Junction.

1.4 Licencing and Purchasing

Unlicensed versions are fully functional and not time-restricted, but all generated TIFF and PDF files have an additional page which indicates that the document was generated with an unlicensed version of the product.

When a license is purchased (see http://www.aquaforest.com/en/buy001.asp) a permanent license key will be issued for use with the product which should be entered using the Help | License option of the GUI. There is no need to download another version of the software.

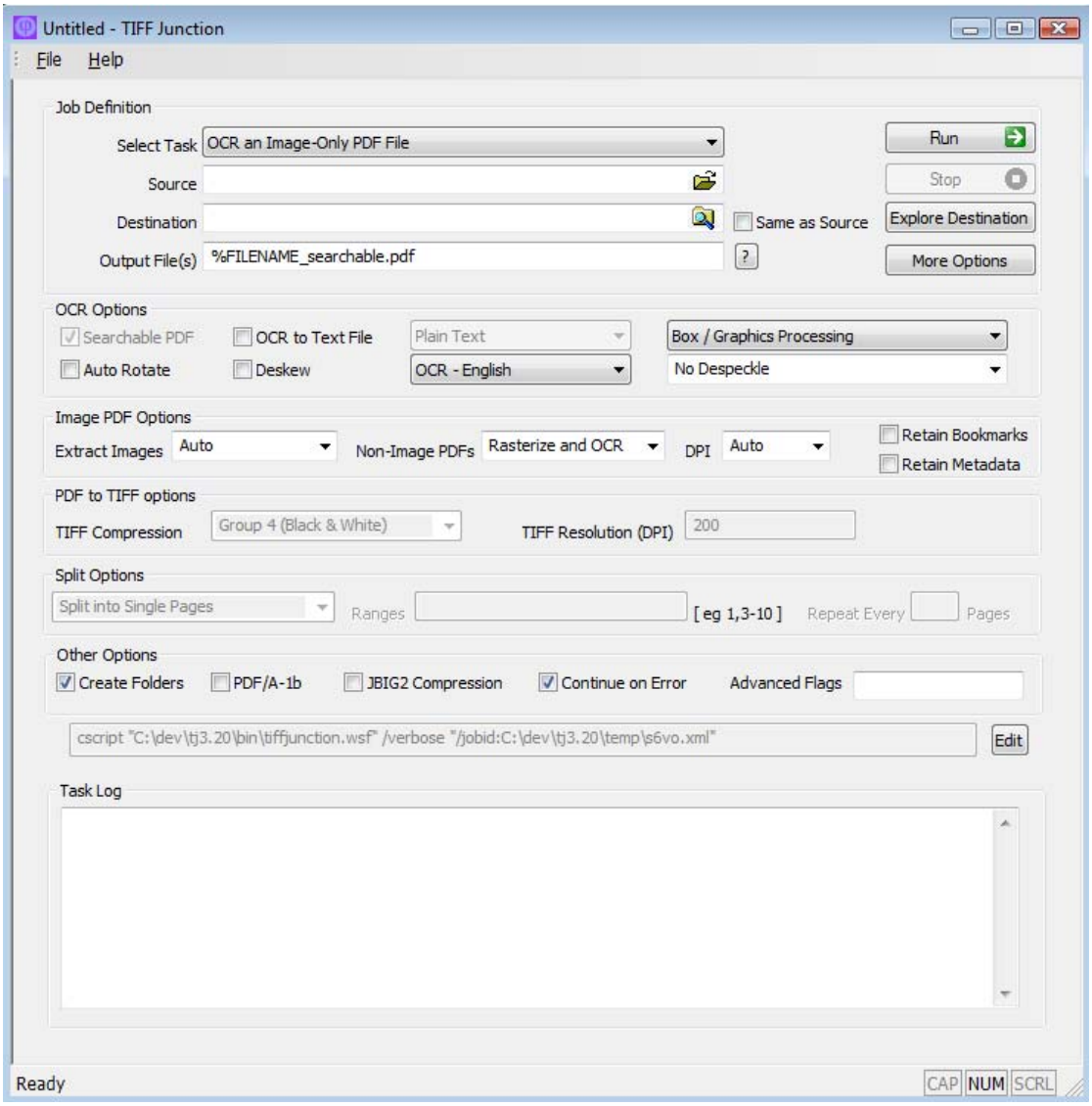
1.5 Conventions

Throughout this guide type in **`bold courier`** is used to represent operating system commands and *TIFFJUNCTIONDIR* represents the directory in which TIFF Junction was installed, by default C:\Progam Files\TIFF Junction.

2 USING THE TIFF JUNCTION WINDOWS INTERFACE

2.1 Main Window

When TIFF Junction is launched, the main window is displayed as shown below. This allows the definition and running of jobs, as well as saving and loading job definitions to and from files.



2.1.1 Job Definition Section

Screen Field / Button	Description
Run	Runs the current job. Job output will appear in the "task log section of the screen".
Stop	Halts processing of the currently running job.
Explore Destination	Launches Windows Explorer using the destination as the starting point.
More Options	This launches a set of property sheets which can be used to set properties of the generated PDF files, such as security and metadata values. See sections 2.2 through 2.5 for more details.

<table>
<tr><th>Screen Field / Button</th><th>Description</th></tr>
<tr><td>Select Task</td><td>This defines which task is to be run and may be one of :

Convert a TIFF file to PDF
Convert a Folder of TIFF Files to PDF
Convert a Folder Tree of TIFF Files to PDF

Merge a Folder of TIFF Files
Merge TIFF files in each Folder in Folder Tree

Merge a Folder of TIFF Files to PDF
Merge TIFF files in each Folder in Folder Tree to PDF

Split a TIFF File
Split a Folder of TIFF Files
Split a Folder Tree of TIFF Files

Show Information for a TIFF File

OCR an Image-Only PDF File
OCR a Folder of Image-Only PDF Files
OCR a Folder Tree of Image-Only PDF Files

Convert a PDF file to TIFF
Convert a Folder of PDF Files to TIFF
Convert a Folder Tree of PDF Files to TIFF</td></tr>
<tr><td>Source</td><td>The source TIFF file or directory.</td></tr>
<tr><td>Destination</td><td>The location where the generated TIFF or PDF file(s) will be placed. This can be set to be the same as the source location.</td></tr>
<tr><td>Output File(s)</td><td>This defines the template output file name.

<table><tr><td>%FILENAME</td><td>Source file name without .pdf</td></tr><tr><td>%UNIQUEn</td><td>For use with split operations. Unique number starting at 1. If n is supplied, then zero padding to n digits is used. If n is not supplied or is zero, then zero padding is not used.</td></tr><tr><td>%DIRNAME</td><td>Source directory name of the directory currently being processed.</td></tr><tr><td>String</td><td>(Any string)</td></tr></table>
For example, a split operation using %FILENAME-part%UNIQUE6.pdf would give filename.pdf split into filename-part000001.pdf filename-part000002.pdf etc.</td></tr>
</table>

2.1.2 OCR Options Section

Screen Field / Button	Description
Searchable PDF (Professional Edition Only)	PDFs generated from TIFF files include searchable hidden text recognised by TIFF Junction's OCR Engine.
OCR to Text File (Professional Edition Only)	Create a separate text file of text recognised by TIFF Junction's OCR Engine.
Other Options (Deskew, Autorotate, Despeckle, Box / Graphics Processing)	See section 5 for further explanation of the OCR options.

2.1.3 Image PDF Options Section

Screen Field / Button	Description
Extract Images	This allows control over the method used to process image PDFs when using the "OCR an Image-Only PDF". The options are : • Auto – TIFF Junction will process the PDF "in-place" without generating a pure image file (rasterization) for processing. • Via Bitmap – The PDF is rasterized using bitmap conversion • Extract TIFF – The embedded TIFF images are extracted directly. • Convert to TIFF – The PDF is rasterized using conversion to PDF. • In Place – TIFF Junction can process the PDF "In-Place" without generating a pure image file (rasterization) for processing. This has the benefit of improved performance in many cases, and where parts of the PDF being processed were "native" (eg converted from Microsoft Word) the "native" portions are kept intact rather than being rasterized, thus voiding the significant size increase that could occur when processing such "mixed" PDF files. Note - Extracting PDF images via TIFF is well suited to PDFs with one scanned image per page, but certain documents that have multiple images on a page, or a mixture of image and text (eg a scanned document with text bates stamps) – for these images the Bitmap or conversion to TIFF method is required.
Non-Image PDFs	This allows control over the treatment of non image-only PDFs, ie PDFs that have some text in them as well as images. The options are : • Rasterize and OCR. The document will be converted to a set of images which will then be processed in the normal way. Note that this actually changes the non-image parts of the PDF to image + text.

	• Raise Error. The task will terminate with an error. If "On Error Continue" is set this then behaves as Skip. This is the default. • Skip. The document will not be processed. • Pass Through. The file will not be processed, but a copy of the document will be made and named as if the processing had occurred.
DPI	Image PDFs are converted to TIFF for OCR processing. This dropdown allows the DPI for the converted TIFF to be specified. If left on "Auto" the DPI will be determined from the original images in the PDF file.
Retain Bookmarks	Bookmarks from the original PDF are copied to the searchable result PDF.
Retain Metadata	Metadata from the original PDF are copied to the searchable result PDF.

2.1.4 PDF to TIFF Options Section

Screen Field / Button	Description
TIFF Compression	For black & white documents "Group 4" should be selected. For color, "LZW" should be used.
TIFF Resolution	This specifies the DPI resolution.

2.1.5 Split Options Section

Screen Field / Button	Description
Split Type	*One of :* Split into single pages Split by ranges (See below) Split by repeating ranges (See below)
Page Ranges	Set of page ranges separated by commas that defines which pages from the original should be extracted. The following types of page ranges are allowed : 1 — Specifies a single page 1-3 — Specifies a range of pages
Repeating Range	Apply the page range to each set of *Page Ranges* pages within the document. For example if 2-4 is specified for page ranges, and 4 is specified as the repeating range, then the range is re-applied every 4 pages. Hence the file is split such that the first output file contains pages 2-4 from the original document, the second contains pages 6-8 and so on.

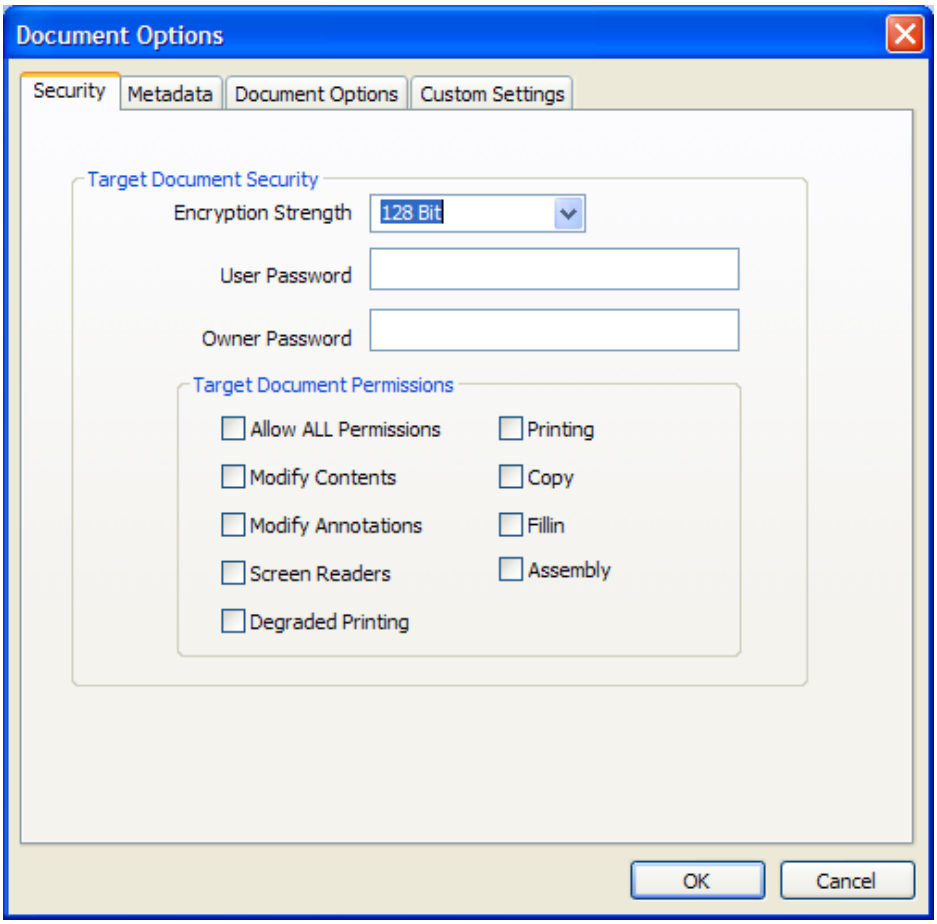
2.1.6 Other Options Section

Screen Field / Button	Description
Create Directories where required	If checked, when processing a tree, subdirectories will be created in the target directory tree when required.
Continue on Error	When checked, folder and tree processing jobs will continue even if there is an error processing individual files.
PDF/A-1b Compliance	Processes the output PDF file to ensure it is compliant with ISO 19005 / PDF/A-1b. See section 6 for more details.
JBIG2 Compression	If this option is chosen, bitonal images in the output PDF will be compressed used JBIG2 compression rather than the default of Group 4. This will result in smaller files for PDFs with black and white images.
Advanced Flags	Additional TIFF Junction advanced command line flags may be entered here (see section 3)

2.1.7 Task Log Section

Screen Field / Button	Description
Command	This is the operating system command that is used to run the job. It can be edited after using the "Edit" button, although there is normally no need for this.
Task Output	The output of the job appears here.

2.2 Security Settings

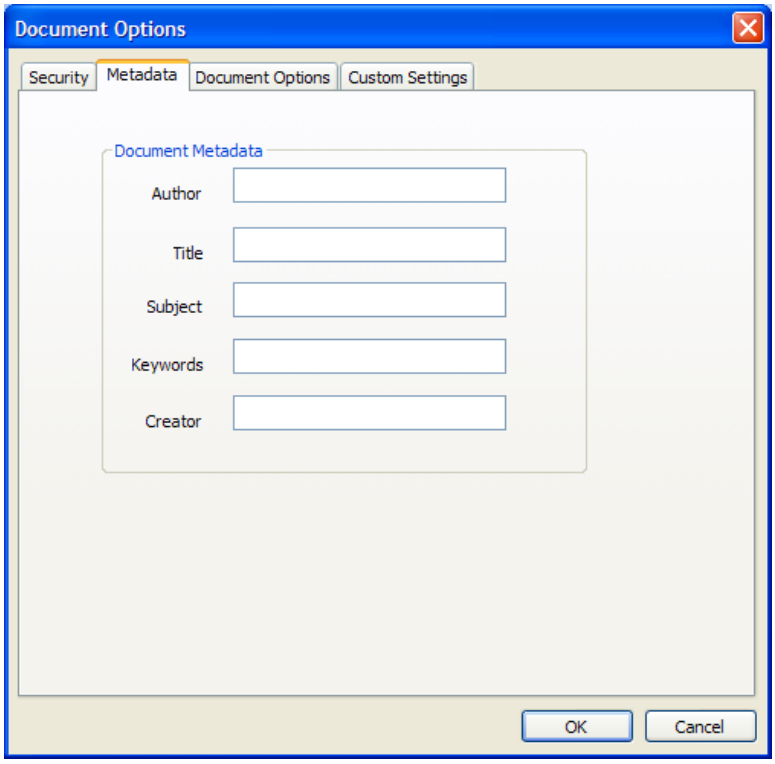


2.2.1 Target Document Security

Option	Description
User Password	A password that will be required to open the document.
Owner Password	A password that will be required to change the document permissions.
Permissions	

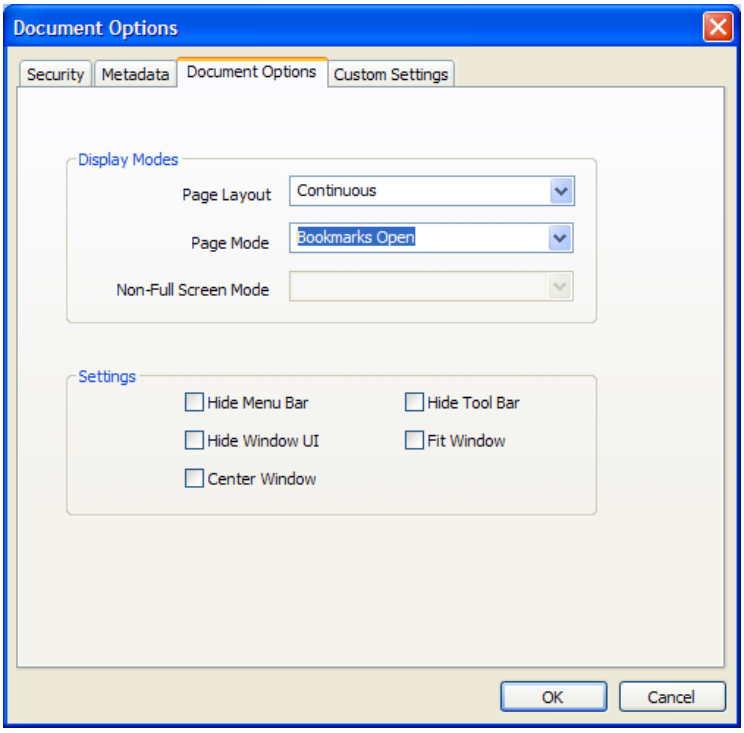
Permission	Description
Allow ALL Permissions	All all the permissions below.
Printing	Allow high-quality printing
Modify Contents	Allow assembly (see below) and other document medications
Copy	Allow text and graphic copying and extraction
Modify Annotations	Allow modification of annotations
Fillin	Allow filling of form fields
Screen Readers	Allow extraction of text and graphics in support of accessibility.
Assembly	Allow rotation, insertion or deletion of pages.
Degraded Printing	Allow low-quality printing

2.3 Document Metadata



This page allows setting of the generated PDF Document Properties.

2.4 Document Display Options



2.4.1 Display Modes

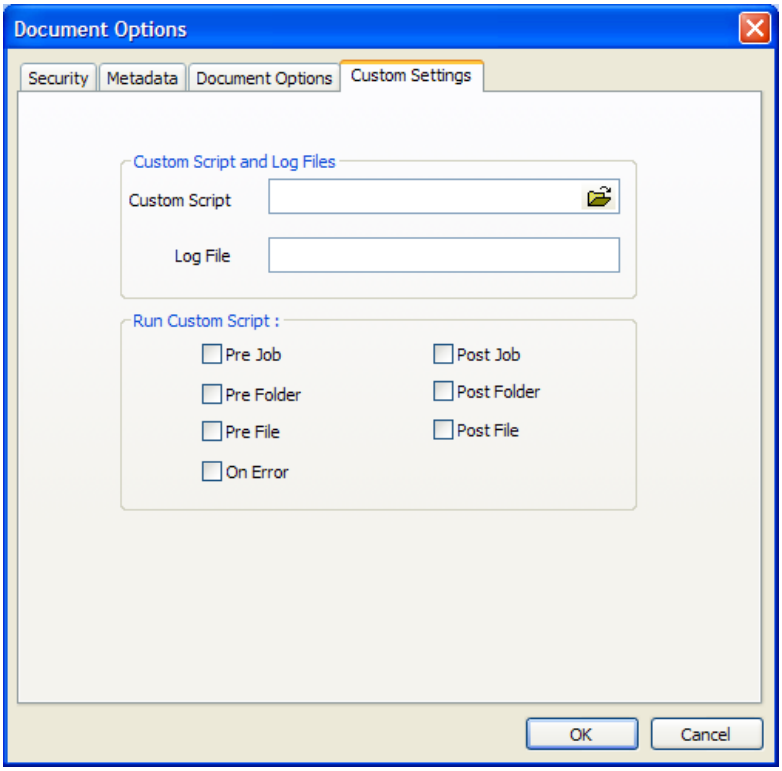
Option	Description
Page Layout	The setting for initial document page display. One of : • Single Page • Continuous • Continuous Facing (odd pages left) • Continuous Facing (odd pages right)
Page Mode	The setting for initial viewer mode. One of : • Neither Bookmarks nor Thumbnails Open • Bookmarks Open • Thumbnails Open • Bookmarks & Thumbnails Open • Full Screen
Non-Full Screen Mode	Only applicable where Page Mode=Full Screen. The setting for document page display when exiting Full Screen mode. One of : • Neither Bookmarks nor Thumbnails Open • Bookmarks Open • Thumbnails Open

2.4.2 Settings

Option	Description
Hide Tool Bar	The viewer's tool bar will be hidden
Hide Menu Bar	The viewer's menu bar will be hidden
Hide Window UI	The viewer's UI elements (scrollbars etc) will be hidden
Fit Window	The viewer will resize the document's window to fit the size of the first displayed page.
Center Window	The document window will be positioned in the center of the screen.

2.5 Custom Settings

See section 5 for full details of running Custom Scripts.



2.5.1 Custom Script and Log Files

Option	Description
Custom Script	The path of a custom script to be used with the Job.

Log File	The name of a log file (which will be placed in *TIFFJUNCTIONDIR*\logs). The log file will contain the same output as is written to the output panel. By default this will contain the same output as is written to the task output. It is possible to configure TIFF Junction to only record errors in the log file by making a modification to the tiffjunction.wsf file (in *TIFFJUNCTIONDIR*\bin) changing
	`var onlyLogErrors=false;`
	to
	`var onlyLogErrors=true;`

2.5.2 Run Custom Script

Option	Description
Pre Job	If checked, the custom script will be called at the start of the job
Piost Job	If checked, the custom script will be called at the end of the job
Pre Folder	If checked, the custom script will be called at the start of processing each folder
Post Folder	If checked, the custom script will be called at the end of processing each folder
Pre File	If checked, the custom script will be called at the start of processing each file
Post File	If checked, the custom script will be called at the end of processing each file
On Error	If checked, the custom script will be called when a processing error occurs.

3 USING TIFF JUNCTION FROM THE COMMAND LINE

The command line usage is shown below

tiffjunction.exe option [-t target] [input file …]

Parameter	Notes
-m	Merge. The set of input files will be merged to form one new Tiff file with the name and location specified by the target parameter.
-s	Split. The input file will by default be split into *n* single page per file Tiff files where n is the number of pages in the input file. Alternate splitting schemes may be specified using the -1 and -2 flags.
-4 *ranges*	Set of page ranges separated by commas that defines which pages from the original should be extracted.
-5 *repatingrange*	Apply the page range to each set of *Page Ranges* pages within the document. For example if 2-4 is specified for page ranges, and 4 is specified as the repeating range, then the range is re-applied every 4 pages.
-p	Convert one or more Tiff files into a pdf file with the name and location specified by the target parameter. If more than one input file is specified, the input file set is first merged into a temporary file and the merged file is converted to pdf.
-i	Provide internal Tiff format information about each page of each input file.
-v	Provide tiffjunction product version and licence key information.
-w	Non image-only PDF file processing directive. See section 2.1.3 for more details. 0 : Rasterize and OCR 1 : Raise Error 2 : Skip 3 : Pass through
-f	Force pass-through of image formats even when Tiff Junction would normally make a conversion to ensure PDF compatability.
-r *resolution*	Force the resolution (dpi) to be passed through as *resolution* instead of the figure held within the file. When processing existing PDF files; if this flag is not specified the file images will be used as a guide : 200DPI or 300 DPI will be used depending upon which is closest to the resolution used in the file.
-d	Force creation of any directories that may be required by the –t parameter.

<table>
<tr><td>-c</td><td>0 : Retain metadata when processing image-only PDFs
1 : Retain bookmarks
2 : Retain bookmarks and metadata

1000 : Force conversion of images to single strip CCITT Group 4 compression, even for images where Tiff Junction would not ordinarily make the conversion. Note – retention of bookmarks and metadata in automatic when using the “In-Place” image extraction of PDF files (see –j below).</td></tr>
<tr><td>-t target</td><td>Target file or directory specification.</td></tr>
<tr><td>-o target</td><td>Target file template.

Output file name specifier. If –f is not specifed, the default parmeter setting is %FILENAME%UNIQUE6% which would give (for example) filename.pdf split into filename000001.pdf filename000002.pdf etc.

<table>
<tr><td>%BOOKMARK</td><td>Bookmark title</td></tr>
<tr><td>%PAGEn</td><td>The first page number from source file, zero padded to n digits. If n is not supplied or is zero, then zero padding is not used</td></tr>
<tr><td>%FILENAME</td><td>Ssource file name without .pdf</td></tr>
<tr><td>%UNIQUEn</td><td>Unique number starting at 1. If n is supplied, then zero padding to n digits is used. If n is not supplied or is zero, then zero padding is not used</td></tr>
<tr><td>%DIRNAME</td><td>Source directory name of the directory currently being processed.</td></tr>
<tr><td>String</td><td>(Any string)</td></tr>
</table>
</td></tr>
<tr><td>-n</td><td>Stamp Name (See Section 9)</td></tr>
<tr><td>-l</td><td>Generate Page Labels using single page file name (See section 10)</td></tr>
<tr><td>-z</td><td>Generate (with -m) or use (-p) Page Labels XML File. (See section 10)</td></tr>
<tr><td>-x</td><td>Stamp Value (See Section 9)</td></tr>
<tr><td>input file</td><td>One or more input file specifications which may include wildcards. Input files are processed in the order that they appear on the command line.</td></tr>
<tr><td>-k option</td><td>Perform OCR processing on the document. Requires the –p flag.

1- Produce Searchable PDF
2- Produce OCR Text File
3 - Produce both Searchable PDF and OCR Text file. With this option the text fle is named filename.txt where filename is the name of the generated PDF file.
4 - Produce Searchable PDF from an Image-Only PDF file.
6 - Produce both Searchable PDF and OCR Text file from an Image-Only PDF file.</td></tr>
</table>

<table>
<tr><td>-g option</td><td>This option controls the type of text document produced by the OCR process.

2 – Plain Ascii (no spacing)
4 – Smart Ascii (space formatting)
6 - HTML (Maximum Quality OCR Engine Only)
8 – RTF (Maximum Quality OCR Engine Only)

If not specified, the default is 4.</td></tr>
<tr><td>-h language</td><td>Sets the dictionary and character recognition to be used for the OCR process :

<table>
<tr><th>LANGUAGE</th><th>-h Flag Value</th></tr>
<tr><td>English</td><td>0</td></tr>
<tr><td>German</td><td>1</td></tr>
<tr><td>French</td><td>2</td></tr>
<tr><td>Russian</td><td>3</td></tr>
<tr><td>Swedish</td><td>4</td></tr>
<tr><td>Spanish</td><td>5</td></tr>
<tr><td>Italian</td><td>6</td></tr>
<tr><td>Russian English</td><td>7</td></tr>
<tr><td>Ukrainian</td><td>8</td></tr>
<tr><td>Serbian</td><td>9</td></tr>
<tr><td>Croatian</td><td>10</td></tr>
<tr><td>Polish</td><td>11</td></tr>
<tr><td>Danish</td><td>12</td></tr>
<tr><td>Portuguese</td><td>13</td></tr>
<tr><td>Dutch</td><td>14</td></tr>
<tr><td>Czech</td><td>15</td></tr>
<tr><td>Roman</td><td>16</td></tr>
<tr><td>Hungar</td><td>17</td></tr>
<tr><td>Bulgar</td><td>18</td></tr>
<tr><td>Slovenian</td><td>19</td></tr>
<tr><td>Latvian</td><td>20</td></tr>
<tr><td>Lithuanian</td><td>21</td></tr>
<tr><td>Estonian</td><td>22</td></tr>
<tr><td>Turkish</td><td>23</td></tr>
</table>
</td></tr>
<tr><td>-1 pdfmetadata</td><td>Sets metadata attributes for the resulting PDF file when using –p.
See section 3.5.1.</td></tr>
<tr><td>-2 pdfoptions</td><td>Sets open options for the resulting PDF file when using –p.
See section 3.5.2.</td></tr>
<tr><td>-3 pdfsecurity</td><td>Sets security attributes for the resulting PDF file when using –p.
See section 3.5.3.</td></tr>
<tr><td>-6</td><td>Deskew (straighten) the image. (Maximum Quality Engine Only)</td></tr>
<tr><td>-7</td><td>Auto-rotate the image – this will ensure all text oriented normally, suitable for OCR (Maximum Quality Engine Only)</td></tr>
<tr><td>-8 pixels</td><td>Despeckle the image – The method removes all disconnected elements within the image that have height or width in pixels less than the specified figure. The maximum value is 9.</td></tr>
<tr><td>-0 [zero]</td><td>When used in combination with the despeckle option, this has the effect of despeckling the page for OCR processing, but the image retained in the final file is not despeckled.</td></tr>
</table>

-j	Defines the method for extracting images from PDFs : 0 – Auto 1 – via Bitmap 2 – Extract TIFFs 3 – Convert to TIFF 4 – In-Place
-a	Ensure that the output file is PDF/A compliant. Note that this cannot be applied to In-Place PDF extraction conversions.
-e	Specifies a temporary folder to be used for bitmap images used during OCR processing. If this is not specified, the first of the following environment variables that is defined will be used : TMP, TMPDIR, TEMP.
-u	Specify that JBIG2 compression should be used for bitonal images.
-9	There are two options that can be used to control how the OCR engine processes parts of the document image that appear to be graphics areas. By default, if an area of the document is indentified as a graphic area then no OCR processing is run on that area. However, certain documents may include areas or boxes that are identified as "graphic" or "picture" areas but that actually do contain useful text. To ensure that the OCR engine can be forced to process such areas there are two options : *"Treat all Graphics Areas as Text"*. This option will ensure the entire document is processed as text. To use this option from the command line use **-9 0** *"Remove Box Lines in OCR Processing"*. This option is ideal for forms where sometimes boxes around text can cause an area to be identified as graphics. This option removes boxes from the temporary copy of the imaged used by the OCR engine. It does not remove boxes from the final image. Technically, this option removes connected elements with a minimum area (by default 100 pixels). To use this option from the command line use **-9 100** (Or replace 100 with a different value >10 if desired). This option is currently only applied for bitonal images.
-q	This command line option should generally only be used under guidance from technical support. It can control the way that color images are processed and force binarization with a particular threshold. (for example -q 127).
-y	Line Removal - This removes lines and boxes during OCR processing to improve recognition – particularly in cases where characters "touch" lines. This option is available via the GUI drop down or via the command line flag –y lr100.5 The values of 100 and 5 are defaults and should only be changed with guidance from Aquaforest technical support. Other advanced Image Morphology options are available using the –y flag . These are rarely required and should be used only under guidance from technical support.
-B	Blank page removal. This option can be used when converting TIFF files to Searchable PDFs. A value should be provided which specifies the pixel threshold to be used to determine whether a page is blank or not. A suggested value is 100 ie using the advanced flag as shown below : -B 100

3.1 Examples of Merging Tiff Files

Merge all the tiff files in d:\aquaforest\tiff\orbera into a file called ace3.tif in the current directory

```
tiffjunction.exe -m -t ace3.tif d:\aquaforest\tiff\orbera\*.tif
```

Merge all the tiff files with a name beginning with "s" in the current directory into a single file called ace2.tif in ..\test

```
tiffjunction.exe -m -t ..\test\ace2.tif s*
```

3.2 Example of Splitting Tiff Files

Create a set of single page per file Tiff files. One file will be created for each page in c:\test\1c00.tif. The output files will be placed in the d:\test directory and named splitaa.tif, splitab.tif etc.

```
tiffjunction.exe -s -t d:\test\split c:\test\1c00.tif
```

Split by ranges : Extract pages 1-3 into one file, and pages 9-12 into another :

```
tiffjunction.exe   -o %FILENAME%UNIQUE6.tif -s -4 "1-3,9-12"  -t
"C:\dev\ts\120f\samples"  "C:\dev\ts\120f\samples\ccitt.tif"
```

Split by repeating ranges, each file will have 3 pages in :

```
tiffjunction.exe   -o %FILENAME%UNIQUE6.tif -s -4 "1-3" -5 3  -t
"C:\dev\ts\120f\samples"  "C:\dev\ts\120f\samples\ccitt.tif"
```

3.3 Examples of Converting Tiff Files to PDF

Convert d:\aquaforest\tiff\709245.tif to a PDF file called ace4.pdf in the current directory :

```
tiffjunction.exe -p -t ace4.pdf d:\aquaforest\tiff\709245.tif
```

Merge all the tiff files with a name beginning with "s" in the current directory into a single PDF file called ace9.pdf in ..\test

```
tiffjunction.exe -p -t ..\test\ace9.tif s*
```

3.4 Creating Searchable PDF Files

Generate a searchable PDF to **`C:\dev\ts\120f\samples\ccitt000000.pdf and a text file to C:\dev\ts\120f\samples\ccitt000000.txt`**

```
tiffjunction.exe   -k 3 -o %FILENAME.pdf –p -t
"C:\dev\ts\120f\samples" "C:\dev\ts\120f\samples\ccitt000000.tif"
```

3.5 PDF Output File Options

3.5.1 metadata specification (-1 Flag)

This may be used to specify metadata properties as in the following example :

```
tiffjunction.exe   -o %FILENAME.pdf -p  -t "C:\dev\ts\120f\samples"
-1 "Author=Tony Blair;Title=Forthcoming Election;"
"C:\dev\ts\120f\samples\ccitt000000.tif"
```

The attributes that may be set are :

- Author
- Title
- Subject
- Keywords
- Creator

3.5.2 docoptions specification (-2 Flag)

This may be used to specify document viewing preferences as in the following example :

```
tiffjunction.exe   -o %FILENAME.pdf -p  -t "C:\dev\ts\120f\samples"  -2
"pagelayoutsinglepage,pagemodeusethumbs,"
"C:\dev\ts\120f\samples\ccitt000000.tif"
```

Propery	**Description**
pagelayoutsinglepage	Single Page (SinglePage)
pagelayoutonecolumn	Continuous(OneColumn)
pagelayouttwocolumnleft	ContinuousFacing [odd pages left] (TwoColumnLeft)
pagelayouttwocolumnright	ContinuousFacing [odd pages right] (TwoColumnRight)
pagemodeusenone	No Thumbnails or Bookmarks Visible (UseNone)
pagemodeuseoutlines	Bookmarks Visible (Use Outlines)
pagemodeusethumbs	Thumbnails Visible (UseThumbs)
pagemodefullscreen	Full Screen Mode (FullScreen)
hidetoolbar	The viewer's tool bar will be hidden
hidemenubar	The viewer's menu bar will be hidden
hidewindowui	The viewer's UI elements (scrollbars etc) will be hidden
fitwindow	The viewer will resize the document's window to fit the size of the first displayed page.
centerwindow	The document window will be positioned in the center of the screen.
nonfullscreenpagemodeusenone	On exiting full screen mode, neither outline or thumbnail images are shown.
nonfullscreenpagemodeusenoutline	On exiting full screen mode outlines are visible.
nonfullscreenpagemodeusethumbs	On exiting full screen mode thumbnails are visible.

3.5.3 security specification (-3 Flag)

This may be used to specify security properties as in the following example :

```
tiffjunction.exe   -o %FILENAME.pdf -p  -t "C:\dev\ts\120f\samples" -3
"strength=128;userpassword=secretsquirrel;permissions=allowdegradedprinting,
"  "C:\dev\ts\120f\samples\ccitt.tif"
```

<table>
<tr><th>Parameter</th><th>Description</th></tr>
<tr><td>Userpassword</td><td>A password that will be required to open the document.</td></tr>
<tr><td>Ownerpassword</td><td>A password that will be required to change the document permissions.</td></tr>
<tr><td>Permissions</td><td>A comma separated list of document permissions
<table>
<tr><th>Permission</th><th>Description</th></tr>
<tr><td>allowprinting</td><td>Allow high-quality printing</td></tr>
<tr><td>allowmodifycontents</td><td>Allow assembly (see below) and other document medications</td></tr>
<tr><td>allowcopy</td><td>Allow text and graphic copying and extraction</td></tr>
<tr><td>allowmodifyannotations</td><td>Allow modification of annotations</td></tr>
<tr><td>allowfillin</td><td>Allow filling of form fields</td></tr>
<tr><td>allowscreenreaders</td><td>Allow extraction of text and graphics in support of accessibility.</td></tr>
<tr><td>allowassembly</td><td>Allow rotation, insertion or deletion of pages.</td></tr>
<tr><td>allowdegradedprinting</td><td>Allow low-quality printing</td></tr>
<tr><td>allowall</td><td>Allow all the permissions</td></tr>
</table>
</td></tr>
<tr><td>strength</td><td>Unless this is set to some other value (such as 40), the default of 128 is used.</td></tr>
</table>

3.6 Directory Processing from the Command Line

Processing of directories requires use of the TIFF Junction WSH (Windows Script Host) file, tiffjunction.wsh. This can be used in conjunction with XML Job Ticket files (See section 4 below) – these can most conveniently be generated by setting the required parameters in the TIFF Junction GUI and using File | Save to save the job ticket file.

Example :

cscript "C:\Program Files\TIFF Junction\bin\tiffjunction.wsf" /verbose "/jobid:C:\dev\tj2\jobdefinition.xml"

Parameter	Notes
/jobid:value	If used, specifies the full path of an XML Job Ticket file saved from the GUI or generated by some other means. Job files in the TIFF Junction jobs directory may be referred to by a shorthand method of *jobname* where the Job Ticket file is *jobname*.xml.
/verbose	Verbose progress messages
/debug	Debug progress messages
/silent	No progress messages
/onerrorcontinue	Allow job to continue after errors

Notes

1. cscript.exe is required to run the command. This is usually included in a standard PATH, but if not would need to be specified explicitly

Eg **`C:\WINDOWS\SYSTEM32\cscript.exe ….`**

2. Options that contain spaces will need to quoted. For example **`“c:\output files”`**

3.7 Converting PDF to TIFF From the Command Line

The command line usage is shown below

```
pdf2tiff.exe  /source=value /target=value /output=value /tiffresolution=value /tiffcompression=value
```

Parameter	Notes
/source	The source PDF file.
/target	The target folder for the output TIFF file
/output	Specifies the output file name. %FILENAME may be used as part of the string and will be replaced by the filename without the file extension. If this is not specified, a value of %FILENAME.tif will be used.
/tiffresolution	The resolution in DPI of the output TIFF file. The default is 200.
/tiffcompression	The image compression scheme to be used in the TIFF file. Either GROUP4 (for bitonal images) or LZW.

4 XML JOB TICKET FILES

TIFF Junction uses XML Job Ticket files to store details of a particular job specification before this is run using the product either from the GUI or using the /jobid option from the command line – see 3.6. Note that some functionality, in particular custom scripts and logfiles can only be used in conjunction with XML Job Ticket files. Other functionality can also be used via the equivalent command line flags.

Developers may wish to create or generate XML job files using the information below. A sample job file is shown here :

```
<pdfjunction_job>
<operation>split</operation>
<sourcetype>folder</sourcetype>
<sourcefiles>C:\qatest\pj\in\set001</sourcefiles>
<target>c:\qatest\pj\out\qa001</target>
<joboptions>-o -c -f %F%\%U%</joboptions>
<metadata></metadata>
<security></security>
<docoptions></docoptions>
<advancedflags></ advancedflags>
<custom_script>custom.wsf</custom_script>
<custom_script_pre_job>true</custom_script_pre_job>
<custom_script_post_job></custom_script_post_job>
<custom_script_pre_folder></custom_script_pre_folder>
<custom_script_post_folder></custom_script_post_folder>
<custom_script_pre_file>true</custom_script_pre_file>
<custom_script_post_file></custom_script_post_file>
<custom_script_on_error></custom_script_on_error>
<logfile>qa001.txt</logfile>
</pdfjunction_job>
```

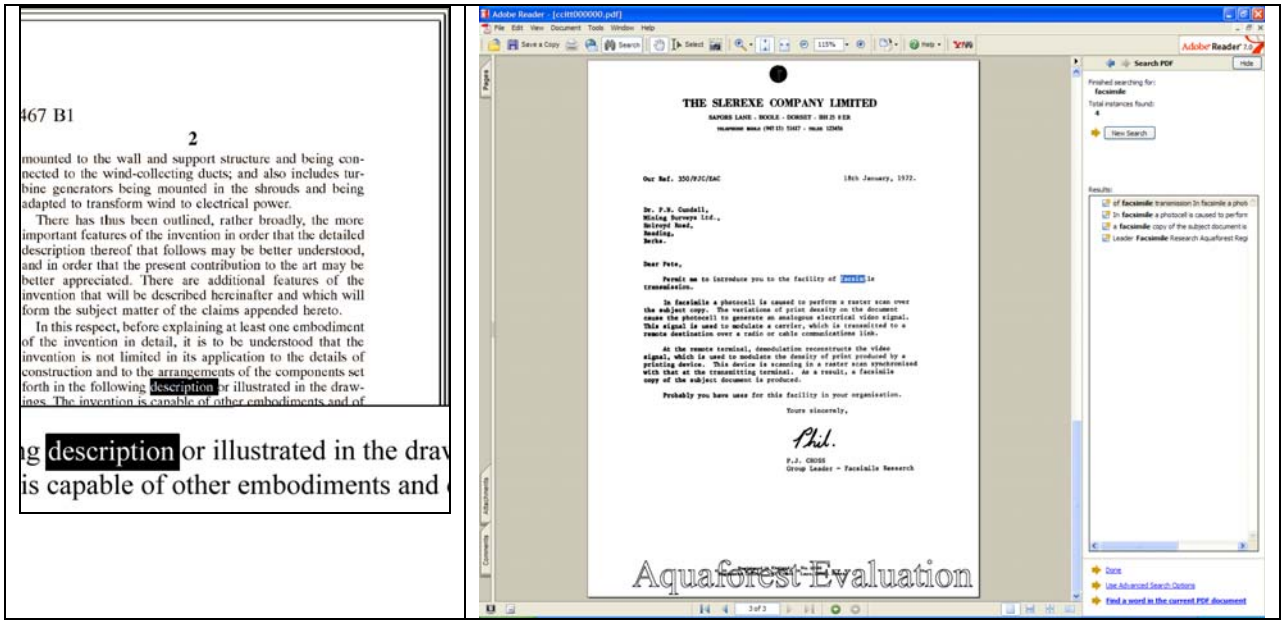
Attribute	Description
operation	Must be split, merge, set or get
sourcetype	Source type (file, folder, tree). The default value is file.
target	Target directory for result files. If not specified, the source directory will be used as the target.
Joboptions	Set of options for the specified operation. Described in detail in section 5.3.
Metadata	Metadata settings for result PDF files. By default metadata fields are left blank. Described in detail in section 5.4.
Security	Security settings for result PDF files. By default files are not secured. Described in detail in section 5.5.
docoptions	Document option settings for result PDF files. Described in detail in section 5.6.
Advancedflags	Advanced command line flags
Sourcefiles	Source TIFF file, folder or tree. May be multiple files for merge operations.
custom_script	If not left blank, specifies the name of a custom windows script file in *PDFJUNCTIONDIR*\custom that will be called in accordance with the custom_script_* settings below.
custom_script_pre_job	If =Y, the custom script will be called at the start of the job
custom_script_post_job	If =Y, the custom script will be called at the end of the job
custom_script_pre_folder	If =Y, the custom script will be called at the start of processing each folder
custom_script_post_folder	If =Y, the custom script will be called at the end of processing each folder
custom_script_pre_file	If =Y, the custom script will be called at the start of processing each file
custom_script_post_file	If =Y, the custom script will be called at the end of processing each file
custom_script_on_error	If =Y, the custom script will be called when a processing error occurs.
logfile	If specified, output will be logged to afile with this name in *TIFFJUNCTIONDIR*\logs

5 CREATING SEARCHABLE PDFS AND TEXT FILES (PROFESSIONAL EDITION ONLY)

By using the Searchable PDF option, TIFFs or Image-Only PDFs are run through an OCR (Optical Character Recognition) process to extract the text from them. The resulting PDF files contain both the images from the original file, togther with a searchable layer of hidden text that may be searched using the standard Adobe Reader search facility as shown below :

5.1 What is a Searchable PDF?

A searchable PDF file is a PDF file that includes text that can be searched upon using the standard Adobe Reader "search" functionality. In addition, the text can be selected and copied from the PDF. Generally, PDF files created from Microsoft Office Word and other documents are by their nature searchable as the source document contains text which is replicated in the PDF, but when creating a PDF from a scanned document and OCR process needs to be applied to recognize the characters within the image.



5.2 Inside a Searchable PDF

In the context of Document Imaging, a searchable PDF will typically contain both the original scanned image plus a separate text layer produced from an OCR process. The text layer is defined in the PDF file as invisible, but can still be selected and seached upon. PDF files are able to store images using most of the native compression schemes used in TIFF files, so for example Group 4 TIFF files do not usually require any format conversion.

5.3 OCR Accuracy

A number of factors affect the accuracy of the text produced by the OCR process – 100% accuracy is certain possible under good conditions but each of the following issues, and OCR processing options will have an impact.

5.3.1 Original Image Quality

Although some pre-processing options such as despeckle and deskew can help in some cases, the visual quality of the original scan is of paramount importance.

5.3.2 Image DPI and Format

The image resolution should be at least 150 DPI for OCR processing, and preferably 300 DPI for optimal results, although for good quality scans 200 DPI is often sufficient. Non-lossy formats (TIFF Group 4, LZW etc) are preferred over lossy formats such as JPEG compression.

5.3.3 Despeckle

This pre-processing option removes isolated “dots” within the image which can cause recognition problems, and makes the result image “cleaner”.

5.3.4 Deskew

This option can improve OCR results by straightening crooked pages.

5.3.5 Auto-Rotate

OCR processing usually recognizes text written top-to-bottom, left-to-right, so pages that are orientated any other way (usually landscape pages) need to be re-oriented to enable recognition.

5.3.6 Speed versus Quality

There are two options that can be used to control how the OCR engine processes parts of the document image that appear to be graphics areas.

By default, if an area of the document is indentified as a graphic area then no OCR processing is run on that area. However, certain documents may include areas or boxes that are identified as “graphic” or “picture” areas but that actually do contain useful text.

To ensure that the OCR engine can be forced to process such areas there are two options :

“Treat all Graphics Areas as Text”. This option will ensure the entire document is processed as text.

“Remove Box Lines in OCR Processing”. This option is ideal for forms where sometimes boxes around text can cause an area to be identified as graphics. This option removes boxes from the temporary copy of the imaged used by the OCR engine. It does not remove boxes from the final image. Technically, this option removes connected elements with a minimum area (by default 100 pixels).

5.3.7 Language Settings

The language setting determines the set of characters that will be recognized, and the dictionary that will be used as a guide.

5.4 The Conversion Process

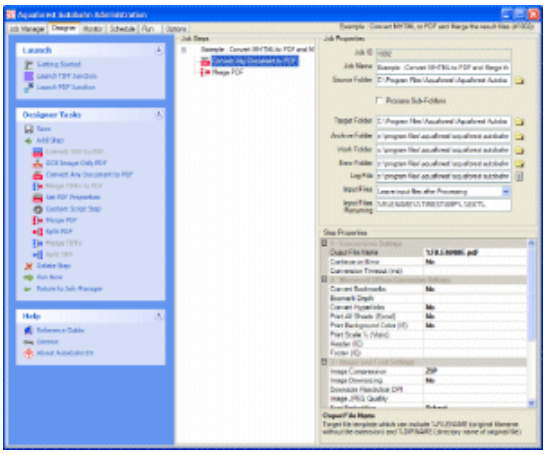
5.4.1 Conversion with TIFF Junction

TIFF Junction is able to convert large volumes of TIFF and Image PDF files to searchable PDF with a high degree of accuracy. When converting from a TIFF file, the process is fairly straightforward; each page image is run through the OCR process according to the options set, and the text layer and image are used to construct the PDF file.

When converting from Image PDF Files, an additional stage is required which creates a TIFF file from the PDF document. By default this is done by rasterizing each PDF page to a bitmap and then converting to TIFF. This ensures a complete representation of each page is made, and is suitable for documents that actually have more than just a single image on each page (for example a Bates number as text) but can be slower than the “image extraction” method which directly extracts the images from each page.

5.4.2 Managing and Scheduling Jobs

Many conversion jobs can benefit from functionality such as Watched Folders, Scheduled Jobs, Windows Service and .Net API. To add these capabilities to TIFF Junction, Autobahn DX is available which includes TIFF Junction as one of it’s components.



5.5 Hardware and Performance

5.5.1 CPU Power

The OCR process is highly CPU intensive and will benefit from being given as much CPU power as possible. As a guide about 1,000 pages per hour can be processed on a 2.5GHz processor, although this will vary according to the source document and OCR options chosen.

5.5.2 Exploiting Multiple CPUs

To take advantage of multiple the multi-threading support and the Job Management facilities of Aquaforest's Autobahn DX product should be used. This includes the same OCR engine as TIFF Junction, but is designed for higher volume, server-based conversions.

5.5.3 Memory

Memory can be a limiting factor when creating the final PDF, in the case of very large documents. A rule of thumb would be to have 1GB – 1.5 GB of memory per processor.

6 CREATING PDF/A COMPLIANT FILES

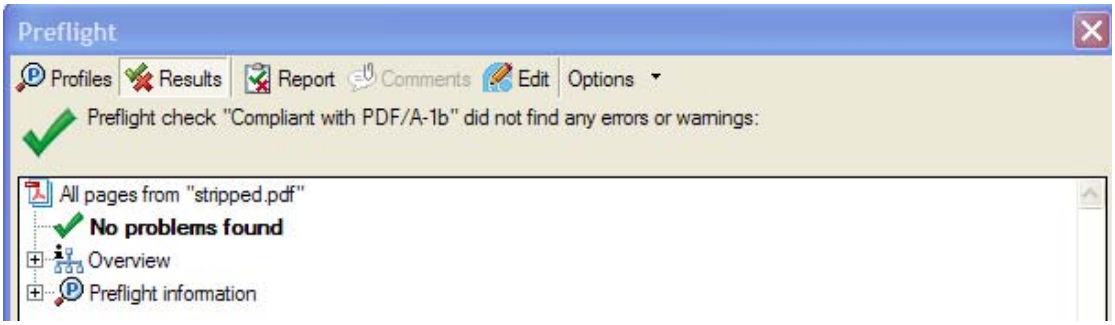
6.1 Background

The PDF/A standard (ISO 19005) defines a file format based on PDF which provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or tendering the files.

There are two levels of conformance possible for PDF/A : Level A and Level B. Documents derived from scanned images are suitable for Level B compliance as they do not include the logical structure, chapters, sections etc required for Level A conformance that may be obtained by (for example) a Microsoft Word document converted to PDF.

6.2 PDF/A Level B Support

TIFF Junction allows generated PDF Files to be generated in a manner that is PDF/A-1b compliant and can be verified as such by using the PDF/A compliance preflight tool in Adobe Acrobat Professional 7.0.7 and later.



6.3 Note : Trial Version

Note that the **trial version** of TIFF Junction places stamps in the PDF files when generating image-only PDF files, and the these stamps do not use embedded fonts and as such will not be PDF/A compliant. When searchable PDF files are generated, the trial stamp is emebedded as part of the image so the resultant files can still be PDF/A compliant if this option is chosen.

6.4 Document Options

The only PDF document settings (from “More Options”) supported in conjunction with creation of PDF/A compliant files are metadata settings for Title, Author, Subject and Keywords. Security options and Document Options are not supported.

6.5 Futher Information

More information relating to ISO 19005-1:2005 and PDF/A can be found on the AIIM website at http://www.aiim.org/standards.asp?ID=25013

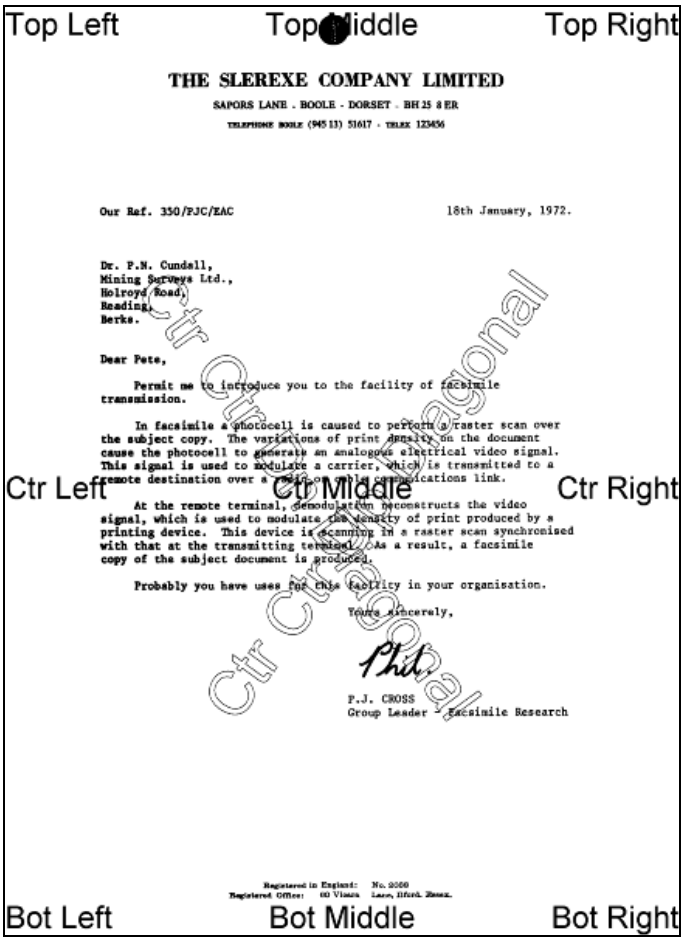
7 USING STAMPS

TIFF Junction allows dynamic and static textual stamps to be placed onto the PDF documents generated by the product by means of a stamps configuration file (stamps.txt in the stamps directory) together with appropriate parameters.

Note : Files generated using Stamps can not be PDF/A compliant in the current release of TIFF Junction.

Note : Stamps are not supported when generating a searchable PDF file. To generate a searchable PDF file with stamps from a TIFF file, first convert to PDF (with the stamp settings) and then use "OCR an Image-Only PDF".

7.1 Stamp Placement



The image above indicates the possible placement for stamps. These are determined according to the parameters in the stamps.txt configuration file. The product comes with a sample set of stamps.

7.2 Stamp Specifications

For example, the stamp that reads "Ctr Ctr Up Diagonal" would be produced by the following lines in the file :

```
BEGINSTAMP
afsName=STAMP
afsText=Ctr Ctr Up Diagonal
afsFontSize=63
afsFont=HELVETICA
afsTextMode=1
afsVpos=CENTER
afsHpos=CENTER
```

```
afsDiag=UP
ENDSTAMP
```

Here is a description of each of the parameters *which are case sensitive :*

Parameter	**Description**
BEGINSTAMP	Required to mark the beginning of a stamp specification.
ENDSTAMP	Required to mark the end of a stamp specification.
afsName	Stamp name. If the stamp name is "STAMP" then the stamp will always be applied. Any other name is used as an identifier and the stamp will only be applied when the –n parameters match the name.
afsText	Stamp text. This may be a fixed piece of text, or may include %p (which will be replaced by the page number) or %s which will be replaced by the value of the -x parameter.
afsFontSize	Point size for the stamp text.
afsFont	Font to be used. The following are supported : TIMES-ROMAN HELVETICA COURIER
afsTextMode	1=Outlined Text 2=Sold Text
afsVpos	Vertical Position of the stamp, which may be one of the following : TOP CENTER BOTTOM
afsHpos	Horizontal Position of the stamp, which may be one of the following : LEFT CENTER RIGHT
afsDiag	Diagonal orientation of the stamp, which may be one of the following : NONE UP DOWN
afsWeight	Font Weight : 0 (default) – 5 (most bold)
afsStartPage	First page of the document to which stamps should be applied (default 1)
afsEndPage	Last page of the document to which stamps should be applied (default 0 which means there is not a limit)
afsPageIncrement	Determines whether stamps should only be applied every *n*th page where n is afsPageIncrement. Default is 0 which means stamps will be applied to all pages (subject to start/end page specifications).

7.3 Example of Using Stamps

Convert 709245.tif to a PDF file called 709245.pdf in the current directory using stamp PRODCODE with a value of **"XR19 Rev1" :**

```
tiffjunction.exe -p –n PRODCODE –x "XR19 Rev1" 709245.tif
```

8 CUSTOM SCRIPTS

The functionality of TIFF Junction can be extended by using custom scripts which may be called at various points during the processing of a particular job. The custom scripts are Windows Script Files which allow scripting in either VBScript or Jscript.

Custom scripts reside in *TIFFJUNCTIONDIR*\custom. In order to implement custom scripts, a developer should take a copy of the script template (custom.wsf , shown below) and implement the script according to their requirements. The custom script name should be specified and required exists should be flagged as being called, either using the GUI or directly in the Job Ticket File (See section 4).

Parameters are passed as follows :

Parameter	**Description**
jobid	The current job id
currentfolder	The path of the folder currently being processed, or “null” for non-applicable exits (eg job_start)
currentfile	The path of the file currently being processed, or “null” for non-applicable exits (eg job_start)
exitpoint	One of the following string values, according to the exit point that was called. pre_job post_job pre_folder post_folder pre_file post_file on_error

```
<job>
<runtime>
  <named name="jobid" helpstring="Job ID" type="string" required="false"/>
  <named  name="currentfolder" helpstring="Current Folder" type="string" required="false" />
  <named  name="currentfile" helpstring="Current File" type="string" required="false" />
  <named  name="exitpoint" helpstring="Script Exit Point" type="string" required="false" />
</runtime>
<script language="VBScript">
'
' Use this section for VBSript
'
'jobID=WScript.Arguments.Named("jobid")
'currentFolder=WScript.Arguments.Named("currentfolder")
'currentFile=WScript.Arguments.Named("currentfile")
'exitPoint=WScript.Arguments.Named("exitpoint")
'WScript.StdOut.WriteLine("Parameter Values - "+jobID+":"+currentFolder+":"+currentFile+":"+exitPoint)
</script>
<script language="JScript">
//
// Use this section for JScript
//
var jobID=WScript.Arguments.Named("jobid");
var currentFolder=WScript.Arguments.Named("currentfolder");
var currentFile=WScript.Arguments.Named("currentfile");
var exitPoint=WScript.Arguments.Named("exitpoint");
WScript.StdOut.WriteLine("Parameter Values - "+jobID+":"+currentFolder+":"+currentFile+":"+exitPoint);
</script>
</job>
```

9 TIFF JUNCTION FILES AND DIRECTORIES

After installation, the following subdirectories are created in the TIFF Junction install directory :

Directory	Contents
bin	Executables and Scripts. OCR resource files
docs	Reference Guide
jobs	Standard directory for XML Job Ticket files.
License	License and License key file
logs	Log files
stamps	Stamps configuration file.
temp	Temporary storage for XML Job Ticket files.

9.1 Temporary Files

The product creates temporary files where necessary in converting from one TIFF format to another – sometimes necessary when generating PDFs to ensure that the TIFF format is compatible with that accepted by PDF.

These files (named TMP999.tif where 999 is the process ID) will be created in the current directory. They will be destroyed upon completion of the process.

10 PDF PAGE LABELS

PDF page labels (which appear under the page thumbnails in Acrobat) by default use the page number. TIFF Junction allows control of these labels in two ways :

10.1 Deriving Page Labels from Source File Names

TIFF Junction allows page labelling to be automatically derived from a file name rather than being sequentially numbered from 1. In the case where a PDF is generated from a set of single page TIFF files, the –l flag may be used. When the flag is set, page labels will be generated by taking the source file name, removing leading zeros and removing the file extension.

10.2 Custom Page Labels

When a multi-page TIFF file is generated from a set of single page TIFF files (using –m), the –z flag allows the generation of a special page label XML file. The file will be placed in the same location as the generated multi-page TIFF (*filename*.tif) and will be named *filename*.xml

The XML file will contain content such as :

```
<page_labels>
<page number=1 label=1></page>
<page number=2 label=1a></page>
…….
<page number=500 label=457></page>
</page_labels>
```

Page labels will be generated by taking the source file name, removing leading zeros and removing the file extension. If required, for any reason, the XML could be editing prior to PDF generation to further cusotmize the page labels.

When using –p to generate a PDF from a multi-page TIFF file (*filename*.tif) , the –z flag will make use of *filename*.xml and generate page labels in accordance with the XML file contents.

11 ACKNOWLEDGEMENTS

This product makes use of a number of Open Source components which are included in binary form. The appropriate acknowledgements and copyright notices are given below.

LEPTONICA
Copyright (C) 2001 Leptonica. All rights reserved.

LIBJPEG
This software is based in part on the work of the Independent JPEG Group.

ZLIB
(C) 1995-2004 Jean-loup Gailly and Mark Adler.

ITEXT 4.1.6
Copyright (C) 1999-2009 by Bruno Lowagie and Paulo Soares et all. All Rights Reserved. Binaries distributed under the Mozilla Public License.

CUNEIFORM

Copyright (c) 1993-2008, Cognitive Technologies. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met

Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution. Neither the name of the Cognitive Technologies nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE

LIBTIFF

Copyright (c) 1988-1997 Sam Leffler. Copyright (c) 1991-1997 Silicon Graphics, Inc.

Permission to use, copy, modify, distribute, and sell this software and its documentation for any purpose is hereby granted without fee, provided that (i) the above copyright notices and this permission notice appear in all copies of the software and related documentation, and (ii) the names of Sam Leffler and Silicon Graphics may not be used in any advertising or publicity relating to the software without the specific, prior written permission of Sam Leffler and Silicon Graphics.

THE SOFTWARE IS PROVIDED "AS-IS" AND WITHOUT WARRANTY OF ANY KIND, EXPRESS, IMPLIED OR OTHERWISE, INCLUDING WITHOUT LIMITATION, ANY WARRANTY OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL SAM LEFFLER OR SILICON GRAPHICS BE LIABLE FOR ANY SPECIAL, INCIDENTAL, INDIRECT OR CONSEQUENTIAL DAMAGES OF ANY KIND, OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER OR NOT ADVISED OF THE POSSIBILITY OF DAMAGE, AND ON ANY THEORY OF LIABILITY, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

12 SUPPORT

Any problems running this application should give evident error messages, but if you are unable to resolve a problem please contact support@aquaforest.com