TIFF Junction 3.2
Release Notes

Version 3.21 - August 2010

http://www.aquaforest.com/

1 Upgrading from earlier Versions

- This release can be installed alongside earlier versions, so it is not essential to uninstall the previous version.

- This release requires version 3.5 of the .NET framework. The setup will check whether this is installed on your system and if not, will take you to the appropriate Microsoft site to download and install .NET 3.5.

- To upgrade from earlier versions request a new license key from Aquaforest : sales@aquaforest.com.

If you have any questions about upgrading to version 3.2 please contact Aquaforest support : support@aquaforest.com

2 Updated OCR Engine

The OCR engine used within TIFF Junction has been changed to the Aquaforest OCR Engine. For the majority of documents this should provide an increase in throughput without any loss of accuracy.

3 New and Changed Job Options

JBIG2 Compression (-u 1 flag)
This option will compress bitonal images in generated PDFs using JBIG2 compression rather than the default Group 4 compression scheme. This will result in smaller PDF file sizes, at a cost of increasing processing time.

DPI Setting (-r flag)
When OCRing a PDF, the PDF is rasterized to produce a TIFF file which is then OCRed. By default the TIFF image resolution is determined from the images embedded in the source PDF but this flag can be used to override default processing and specify the DPI of the TIFF that will be generated.

Box / Graphics Options (-9 flag).
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 (Note : this replaces the previously undocumented -9 option)

“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 if desired). This option is currently only applied for bitonal images.

Despeckle (-8 flag).
A new despeckle algorithm has been incorporated in this release. The maximum despeckle value is now 9. Higher values passed via the -8 flag will be reduced to 9. The method removes all disconnected elements within the image that have height or width in pixels less than the specified figure.

OCRQuality (–y flag).
This speed versus quality does not apply in the new OCR engine. The –y command line flag for this purpose has been reassigned to Image Morphology (see below)

OCR Languages
There have been a number of changes to the list of supported languages and the respective –h flag values as shown below.

LANGUAGE	-h Flag Value
English	0
German	1
French	2
Russian	3
Swedish	4
Spanish	5
Italian	6
Russian English	7
Ukrainian	8
Serbian	9
Croatian	10
Polish	11
Danish	12
Portuguese	13
Dutch	14
Czech	15
Roman	16
Hungar	17
Bulgar	18
Slovenian	19
Latvian	20
Lithuanian	21
Estonian	22
Turkish	23

Binarization (-q flag)
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).

Image Morphology (-y flag)
This command line option should generally only be used under guidance from technical support (except for Line Removal – see 7 below).

4 COMMAND LINE DISPLAY

To assist developers wishing to use the command line interface the TIFF Junction task log portion of the GUI will display the underlying tiffjunction.exe parameters used when a job is run.

Note that the TIFF Junction command line cannot be used to automate processing of multiple documents in a batch or server environment as this is prohibited by the license. Autobahn DX Server should be used to meet this requirement.

5 IMPROVED IMAGE PDF PROCESSING (FROM VERSION 3.21)

An enhanced method for processing Image PDFs is now available as a new option in addition to the existing “Auto”, “Via Bitmap” and “Convert to TIFF” options. 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 avoiding the significant size increase that could occur when processing such “mixed” PDF files.

6 BLANK PAGE REMOVAL (FROM VERSION 3.21)

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. If a page is deemed to be blank then it is omitted from the output file. A suggested value is 100 ie using the new –B advanced flag : -B 100

7 LINE REMOVAL (FROM VERSION 3.21)

A new option to “Remove Lines in OCR Processing” is available. 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.

8 DOT MATRIX IMAGE PROCESSING (FROM VERSION 3.21)

A new option – the –D flag – is available to produce optimized results for dot matrix printed documents. This option will significantly improve recognition for dot matrix documents but should not be used when processing other document types as it will have a negative impact on non-dot matrix documents.