Autobahn DX 2.20 Release Notes

August 2009

http://www.aquaforest.com/

1 Upgrading from earlier Versions

- 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 2.20 please contact Aquaforest support : support@aquaforest.com

2 Updated OCR Engine

The OCR engine used within Autobahn DX has been changed to the Aquaforest OCR Engine. For the majority of documents this should provide an increase in throughput with comparable OCR accuracy.

3 New and Changed Job Options

JBIG2 Compression
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
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
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).

Despeckle 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
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.

4 ADDITIONAL CUSTOM STEP SAMPLES

Section 7 of the reference guide now includes samples to convert postscript to PDF and .msg files to PDF.