Autobahn DX 5.0
Release Notes

Version 5.0
August 2019

Web: www.aquaforest.com
E-mail: info@aquaforest.com

Aquaforest

1 VERSION 5.0.190805

1.1 Bug Fixes

1.1.1 SharePoint connector license check failure

Ref: ADX-222

The previous release of Autobahn DX 5.0 failed executing the SharePoint steps, this was because of a bug in the license key validation

2 VERSION 5.0.190715

2.1 Bug Fixes

2.1.1 Any File to PDF fails when generating PDF/A files from text documents

Ref: ADX-218

The previous release of Autobahn DX 5.0 failed when generating PDF/A files from text documents when using **GenericExtension, AutoExtension** and **AutoExtensionEx.** Updating to the latest version of **BCL** fixes this issue.

2.1.2 Batch Size gets dropped after the first iteration of job from the service

Ref: ADX-220

The previous release of Autobahn DX 5.0 was setting the job filter limit to zero after the job executes for the first time under the service.

3 VERSION 5.0.190605

3.1 Bug Fixes

3.1.1 CSV Log Files for Paths with Comma(s)

In the previous version of Autobanh DX, the presence of commas (,) in file paths adds unwanted columns in the CSV log file. We have fixed this issue.

3.1.2 Merging PDF Files with Acroforms

In the previous version of Autobanh DX, there was a bug that was causing merging of PDF files with Acroforms to fail. We have fixed this issue.

4 VERSION 5.0.190430

4.1 Upgrading from earlier Versions

- This release requires version 4.5.2 of the .NET framework. The setup will check whether they are installed on your system and if not, will take you to the appropriate Microsoft site to download and install.

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

- Upgrade blog: http://www.aquaforest.com/wp/index.php/upgrading-autobahn-dx-server/

4.1.1 Preserving Existing Job Definitions when Upgrading

When Upgrading to a new version of Autobahn DX, your old jobs will not have all the new step properties added. To rectify this issue, open all your old jobs from the **Job Manager** and save them.

4.1.2 License Key

Autobahn DX 5.0 uses different license keys from the previous versions of Autobahn DX. You will need to request a new license key from Aquaforest: sales@aquaforest.com.

4.2 Enhancements in v5.0

We have made a lot of changes in this version of Autobahn DX, we will discuss these enhancements in this section.

4.2.1 Pause Job

We have now added the ability to resume from Jobs in Autobahn DX if:

1. The Job is Interrupted By a service crash or power failure.
2. If you paused the job from the Autobahn DX GUI.

Note: If you make any changes to the Job when it is in a Paused state the job will start from the beginning.

4.2.2 New Job Steps

In Autobahn DX 5.0, we have added to our long list of job steps. This is to give the user more value and options. For more details, check the section 5.7.2 in the Autobahn DX 5.0 reference guide.

4.2.2.1 Cloud OCR

The optional Cloud OCR module extends Autobahn DX with additional OCR engines from **Microsoft** and **Google**, the main advantages of these OCR engines is their Hand writing

recognition capabilities. These OCR engines are available as a SAAS model provided by both vendors. Before you can start using these steps in Autobahn DX, you will need to have a subscription first. See chapter 18 of the reference guide for more details.

We have added two step types to the **Advanced** section of the **Job Designer** tab of Autobahn DX, the steps are named:

- Image to Searchable PDF (Cloud OCR)
- PDF to Searchable PDF (Cloud OCR)

4.2.2.2 Stamp PDF Files

This step can be used to add stamps to PDF pages, we have given the user the ability to customize these stamps extensively in a very simple manner.

Autobahn DX has different ways to apply stamps to a page, this gives the user some level of flexibility.

- **StampTextAsString**: When this operation has selected the text passed as the **StampObject** will be stamped on the PDF document as text.
- **StampPDFText**: When this operation is selected the text passed as the **StampObject** will be stamped on the PDF document as an image.
- **StampPageNumber**: When this operation is selected, every page in the PDF file will be stamped with a page number, starting from the start number. E.g. if StartNumber = 6 the first-page number will start from 6.
- **StampPageNumberBates**: When this operation is selected, every page in the PDF file will be stamped with a bate number, starting from the start number. E.g. if **StartNumber** = 6 the first-page number will start from 000006.
- **StampVariable**: This option allows a user to specify a variable like a date, filename or time. The variable specified by the **StampObject** will be stamped on the document. Check the table below for different Stamp variables provided.
- **StampPDFImage**: When this operation is selected the text passed as the **StampObject** is the address of the image to be stamped on the PDF document.

4.2.2.3 Any File to Searchable PDF (Extended)

In previous versions of Autobahn DX, we use to have the **OCR Any File to PDF** (this has changed to **Any File to Searchable PDF (Standard)**) step. This step converted office files to PDF and performed an OCR on image-based files. This step use to be available only for the Standard OCR engine, in version 5.0 we have added similar step that will use the Extended Engine to OCR image-based files.

4.2.2.4 Azure Storage Download

We added this new step to allow users to download files from an Azure Storage Container to your local machine. This can be used as part of a workflow in Autobahn DX.

4.2.2.5 Azure Storage Upload

We added this new step to allow users to upload files to an Azure Storage Container from your local machine. This can be used as part of a workflow in Autobahn DX.
Using these two steps, you can download files from Azure, process them and upload the outputs back to Azure in a single job.

4.2.2.6 Distributed Polling

This step can be used to implement load balancing in Autobahn DX, it achieves this by copying a fraction of the files from a central input location to the local system where Autobahn DX is running. Multiple Autobahn DX servers can point to one input folder, as a result, the files will be shared across several servers and the processing will be more optimized.

4.2.3 Job API Changes

4.2.3.1 Remote API Enhancement

Previously you had to install Autobahn DX on the client and server machine in other to call a remote API in Autobahn DX. We have changed this so that you will only need to install Autobahn DX on the server computer.

4.2.3.2 GetLastRunDate

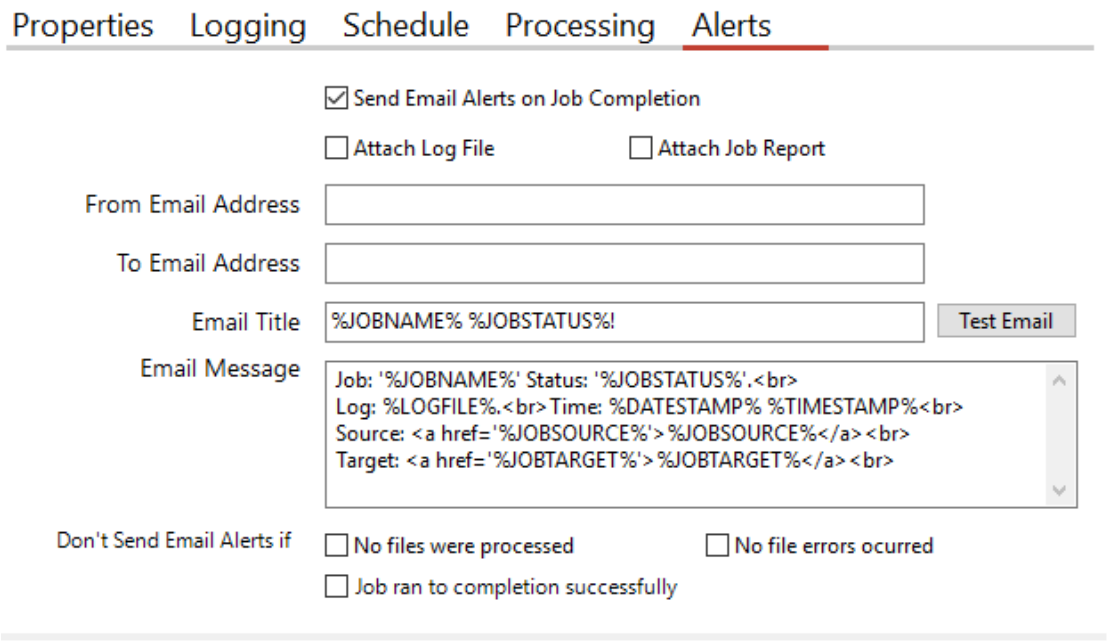
We have added the method below to the Job API
Public string GetLastRunDate();
Returns the last Date and Time the job executed.

4.2.4 New Alerts Method

We have changed the way alerts are setup to give the user more control over when to send alerts and what to include in the alerts.

Note: If you are upgrading your jobs from a previous version of Autobahn DX and you have alerts setup for the job, you will have to go the **Alerts** tab in the **Job Designer** and set up the alerts in your jobs again.

See section 5.2.4 of the reference guide for more details.



4.2.5 OCR Updates

4.2.5.1 Extended Engine

Autobahn DX 5.0 now has the latest version of the iDRS engine (iDRS 15.4.2) in the Extended OCR module.

4.2.5.1.1 Default Values

The default values for a few settings have been changed so that it gives good OCR results for different types of documents. These are shown below:

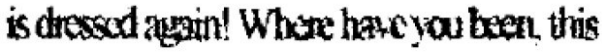
Setting	Changed to
Binarize	true
Binarization Mode	Adaptive
Brightness	128
Smoothing Level	248
Threshold	0
Work Depth	255
Remove Lines	true

4.2.5.1.2 New High Quality OCR engine

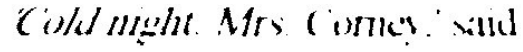
The iDRS™ is updated with I.R.I.S.' brand-new High-Quality OCR: a new OCR engine developed using state of the art concepts from the artificial intelligence research domain.

This new technology brings considerable OCR accuracy improvement especially for bad quality scans, camera images or low-resolution documents, which are affected by common issues such as:

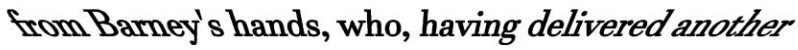
- **Touching characters**



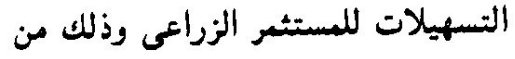
- **Broken characters**



- **Distorted characters**



It will also be suited for recognition of Arabic and Farsi, due to the cursive nature of these languages:



The first release uses High Quality OCR engine for English, Arabic and Farsi languages; further languages will of course be added in future releases.

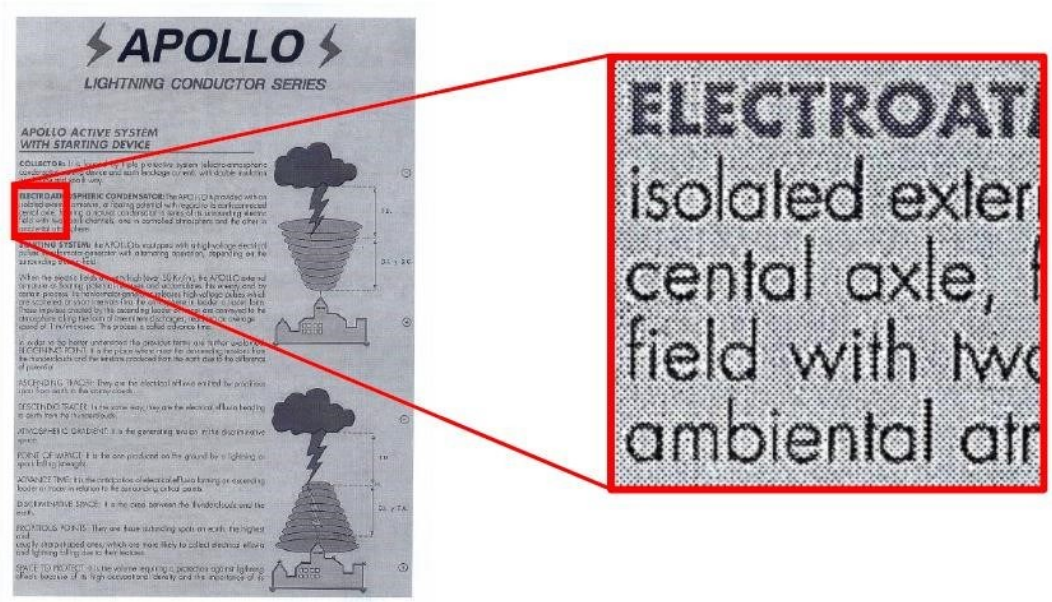
- For Latin, Cyrillic, Greek, Hebrew and Asian languages, High Quality OCR will be combined with existing OCR engine to use the strengths of both engines.

- For Arabic and Farsi languages, it fully replaces the previous engine, and reaches an unparalleled level of accuracy.

Note that processing time with High Quality OCR engine is expected to increase for low-quality documents: more time will be spent but better accuracy will be reached.

4.2.5.1.3 Recognition of images scanned with dithering

This release exposes an option allowing to improve recognition of color or greyscale images scanned with dithering:



Previous releases would not have properly processed such images: in most cases, the text would simply not have been detected during page analysis step.

How to use

It can be enabled by setting the Undithering property in the Binarization object. Note that you also need to enable smoothing by setting SmoothingLevel to a value greater than '0' to perform undithering.

4.2.5.1.4 Automatic language detection of a single-language page

Extended OCR can now automatically detect the language of an input document.
The aim of this feature is to detect the most probable language of a single-language page.

Supported languages

This release will be able to reliably detect the following scripts/languages:

- **Latin script**

  English, German, French, Spanish, Italian, Swedish, Danish, Norwegian, Dutch, Portuguese, Galician, Icelandic, Czech, Hungarian, Polish, Romanian, Slovak, Croatian, Slovenian, Finnish,

  Turkish, Estonian, Lithuanian, Latvian, Albanian, Catalan, Irish Gaelic, Scottish Gaelic, Basque, Indonesian, Malay, Swahili, Tagalog, Haitian Creole, Kurdish, Cebuano, Ganda, Kinyarwanda, Malagasy, Maltese, Nyanja, Sotho, Sundanese, Welsh, Javanese, Azeri (Latin), Uzbek, Bosnian (Latin), Afrikaans
- **Cyrillic script**

  Serbian, Russian, Byelorussian, Ukrainian, Macedonian, Bulgarian, Kazakh
- **Greek script**

  Greek
- **Hebrew script**

  Hebrew

Future releases will extend the support to Arabic and Asian scripts.

Note:

- If at least one language has been detected, recognition will be performed in the first language candidate that has been detected, and not in the language(s) set through the **OCR Language x** property.
- If it fails to detect a language, recognition will be performed using the language(s) set through the **OCR Language x** property.

4.2.5.1.5 Punch-hole removal

A new feature has been added to the Extended engine that attempts to remove punch holes from pages. This feature only works when converting images to PDFs or when OCRing PDFs with **Extract Images Method** set to **Convert to TIFF** and with either **Keep Original Image** set to false or **Keep Punch Hole Removal** set to true.

Note: The punch-hole algorithm can be used on images with the following minimum dimensions width: 300px, height: 100px (computed for 300 DPI). The minimum height and width can vary with the image resolution.

4.2.5.1.6 Retain pre-processing settings

You can now retain specific pre-processing in the output PDF documents. For instance, if de-speckling is enabled, speckles are removed from each page to improve the OCR recognition, but this is only done internally and are not reflected in the output PDF document.

In this release, if you want to retain the de-speckling in the output document, set **Keep Despeckled Image** to true. Other pre-processing settings that can be preserved are **deskew**, **dark border removal** and **punch-hole removal**. These can be enabled using **Keep Deskewed Image**, **Keep Dark Border Removal** and Keep **Punch Hole Removal** respectively.

This feature only works when converting images to PDFs or when OCRing PDFs with **Extract Image Method** set to **Convert to TIFF** and with **Keep Original Image** set to false.

4.2.5.1.7 Advanced pre-processing settings

This release has new advanced settings for some existing pre-processing settings of the Extended module. These are:

- AdvancedDeskew
- AdjustmentMode
- ForceDeskew
- AdvancedDespeckle
- Dilate

4.2.5.1.8 New languages available with High-Quality OCR engine

The brand-new technology 'High-Quality OCR' now embeds the 3 following languages:

- Italian
- Spanish
- Portuguese

Note also that variants of already existing High-Quality OCR languages are now supported as well: Afrikaans, Brazilian Portuguese, British, Corsican, Frisian, Luxembourgish, Mexican Spanish, Sardinian, and Swiss-German.

4.2.5.1.9 Performance improved for page orientation detection on Korean documents

The algorithm used for page orientation detection with Korean language has been reviewed, allowing to drastically reduce processing time while improving a bit the accuracy.

On a set of 132 Korean documents, taken in all possible orientations for a total of 528 test cases:

- Older versions:
  - Total time for orientation detection: 5,864 seconds
  - Orientation detection accuracy: 96,0%
- This version:
  - Total time for orientation detection: 971 seconds (divided by a factor 6!)
  - Orientation detection accuracy: 97,3%

4.2.5.1.10 Memory consumption reduced for document conversion

The document output engine includes several optimizations regarding memory consumption when creating an output document. Those changes impact mostly the creation of PDF Image-Text and especially PDF iHQC documents.

In terms of peak memory consumption, considering an input image A4 at 600DPI:

- Older versions:
  - PDF Image-Text: 343 Mb
  - PDF iHQC: 568 Mb
- This version:
  - PDF Image-Text: 238 Mb
  - PDF iHQC: 359 Mb

4.2.5.1.11 Turn off PDF/A validation

In previous versions, PDF/A validation was always performed after converting to PDF/A. However, validating a PDF/A document adds a small performance penalty in terms of the overall processing time. This version allows you to turn off PDF/A validation.

4.2.5.1 Standard Engine

4.2.5.1.1 Default Values

The default values for a few settings have been changed so that it gives good OCR results for different types of documents. These are shown below:

Setting	Changed to
SavePreDespeckle	true

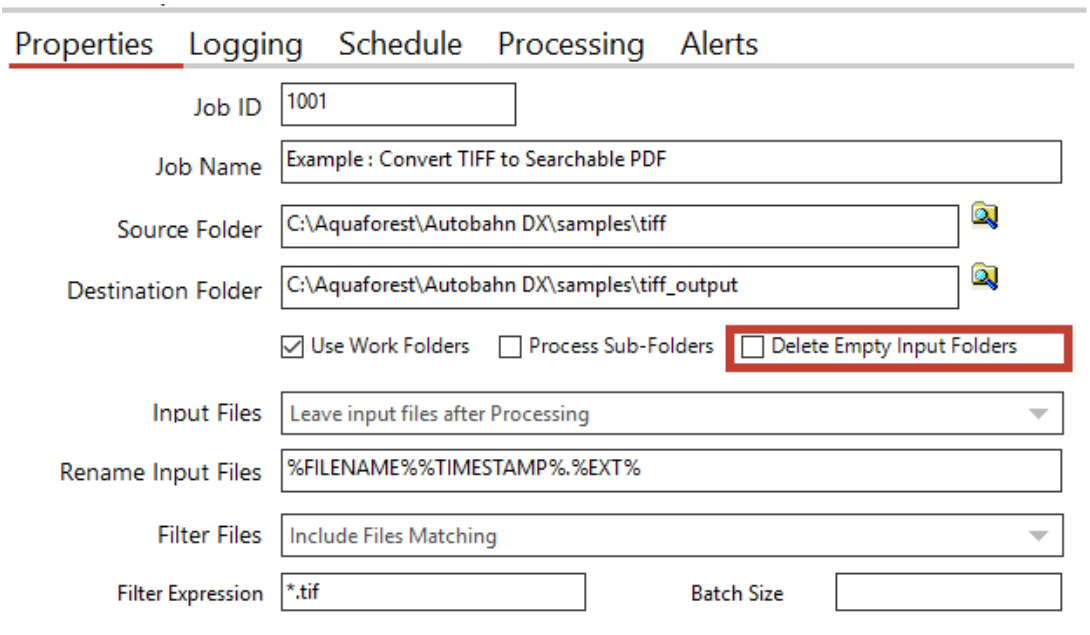
4.2.6 Step Types that have changed name

For clarity we have changed the names and groupings of our OCR steps in Autobahn DX to more clearly represent what they do. The table below shows the old step names and the corresponding new step.

Old Step Name	New Step Name
Convert TIFF to PDF	Image To Searchable PDF (Standard)
Extended Convert TIFF to PDF	Image To Searchable PDF (Extended)
OCR Image-Only PDF	PDF to Searchable PDF (Standard)
Extended OCR Image PDF	PDF to Searchable PDF (Extended)
OCR Any File to PDF	Any File to Searchable PDF (Standard)
Merge TIFFs to PDF	Merge Image to Searchable PDF (Standard)
Extended Merge TIFF to PDF	Merge Image to Searchable PDF (Extended)

4.2.7 Delete Empty Input Folders

When users select **Delete Input Files** or **Move to Archive after Processing** as the input file post processing action, it is a usual occurrence for a lot of empty folders in the input folder tree to remain. To delete these empty folders, you can use this new setting provided in Autobahn DX 5.0.



4.2.8 CPU Core Licensing and Job Control

Your license key will support a specific number of CPU cores. The product will limit the number of concurrent file processing operations to this number and will "throttle" jobs accordingly.

For example, if a 4 core licensed server is currently running a 2 core job and a new job starts that is configured for 4 cores the number of cores allocated to the second job will be reduced accordingly:

Autobahn DX using 2 cores out of 4 allowed.
We will reduce the number of cores in this job from 4 to 2 allowed.

As another example, if a 4 core licensed server is currently running a 4 core job and a new job starts that is configured for 2 cores then the second job will not be able to start until cores are freed up:

Autobahn DX using 4 processors out of 4 allowed.
We will attempt to start the job 18 time(s) over the next 180 seconds.

The retry interval and number of tries is determined by these two config file settings in Autobahn.config (by default this file is in C:\Aquaforest\Autobahn DX\config)

<add key="jobqueuetimeout" value="180" />
<add key="jobqueueinterval" value="10"/>

4.2.9 Autobahn DX Directory Changes

We have added a distribution directory to the installation directory of Autobahn DX, this directory will contain the components need for Autobahn DX to function. As a result, we have moved some folders from the top-level folder to the distribution folder, we have also created new folders for other components. The table below shows the details.

Application	**Old Directory Path**	**New Directory Path**
Extended OCR	extendedocr	distribution /extendedocr
TIFF Junction	tj	distribution /tj
PDF Junction	pj	distribution /pj
Cloud OCR (new)		distribution /cloudocr
SharePoint Connector (new)		distribution /sharepoint
Azure Connector (new)		distribution /azure
Support Tool	support	distribution /support

4.3 Bug Fixes

4.3.1 [SDK-120] Graphics state

The graphics state was not being restored when processing pages that require rotation in the Standard OCR engine. This caused issues when other applications manipulated the PDF after it had been OCRed by Aquaforest. This has now been fixed.

4.4 Known Issues

4.4.1 Recognition of accented characters with High-Quality OCR engine (Extended OCR module)

The new Extended OCR module currently has an issue that impacts Latin languages processed with HighQuality OCR engine.

When a character with an accent (like é, è, à, ñ, etc.) is recognized but is not present in the character set (for instance if recognition is performed in English), the OCR engine will output a reject character (U+FFFD).
This is a regression compared to previous versions, where the "base" character would be output instead (e.g. ‘e’ instead of ‘é’).

This issue will be fixed with the next release.