Aquaforest

Aquaforest SDK
Release Notes

Aquaforest

Making Content Findable

Aquaforest SDK

Release Notes

Version 3.0

October 2020

1 Version 3.0

1.1 New Features

- SDK 3.0 has the following new components
    - Data Extraction engine to automatically extract name-value pairs from PDF documents
    - PDF Toolkit to manipulate PDF files
    - Cloud OCR engine that supports OCRing documents using the Google or the Microsoft OCR engines
- Added a new extensible logger. Previous versions only supported outputting log information to the console or a file. A new `IAquaforestLogger` interface has now been added that can be extended to output anywhere you want. It also has options to set the log level to either Debug, Information, Warning, Error to filter out logs. The logger can be injected into the constructor of all SDK 3.0 components. As a result of this change, `EnableConsoleOuput` and `EnableDebugOutput` settings have been removed as they can now be controlled by the logger object. A reference to `Aquforest.Logging` must be added to use the `IAquaforestLogger` interface

1.1.1 OCR

- [SDK-143] Add new iDRS engine in Extended OCR
    - PDF/A-3a and PDF/A-3b output is now supported
    - Vietnamese and Thai language is now supported
    - Added new `BlankPageDetectionMode` property that allows choosing which algorithm to use for blank page detection
- [SDK-151] - Add options to stop processing when page(s) fail(s) to process in Native mode. Previously, if the OCR engines failed to OCR all pages of a document in Native, the original source document was outputted. However, this made it difficult to mark and identify the document as failed. The new `NativeProcessingErrorMode` property can be used to indicate to the OCR engine how to respond if there are issues processing page(s) in Native mode. The available options are:
    - Do not stop processing if a page fails to process
    - Stop processing only if all pages fail to process
    - Stop processing if at least one page fails to process
- [SDK-152] - Add a method or property to check if PDF is secured
- [SDK-153] - Implement mobile capture extension in Extended OCR. This is controlled by the `PerpectiveCorrection` setting

1.1.2 PDF Toolkit

- [PTK-2] - Enable access to co-ordinates for words in extracted text
- [PTK-8] - Ability to create/retain bookmarks
- [PTK-15] – Extract Text from areas
- [PTK-23] - Add the ability to flatten PDF Form (XFA Forms)
- [PTK-27] - Image To PDF Converter
- Read XFA form data

1.2 Improvements

1.2.1 OCR

- Updated PDF tool used in the Extended OCR engine to process PDFs in Native mode
- [SDK-128] - Check validity of "Temp" Folder to provide additional debug information to customers
- [SDK-130] - Implement iDRS 15.4.5 OR higher to address using ForceTableZones in conjunction with Arabic Language
- [SDK-155] Improve memory usage when processing large files
- [SDK-123] Changed the way the Extended OCR engine is initialised. Previously, the engine was initalised when a new `Ocr` object was created.

```
using (Ocr ocr = new Ocr(resourceFolder))
{
    […]
}
```

Now it needs to be initialised on application start and unloaded at the end.

```
ExtendedOcrEngine.SetupOcr(LICENSE, RESOURCES_FOLDER);

using (Ocr ocr = new Ocr(logger))
{
    […]
}

ExtendedOcrEngine.UnloadOcr();
```

1.2.2 PDF Toolkit

1.3 Changes

- SDK 3.0 is built against .NET Framework 4.7.2
- The license key, resource folder and an optional logger must now be specified in the constructor of OCR, Barcode and Data Extraction engines. As a result, `License` and `ResourceFolder` properties have been removed from these products
- `DecodeResult.BarcodeResult` has been removed from the Barcode engine. To access barcode results, use `DecodeResult.BarcodeResults` instead
- [SDK-129] Removed `EmbedFonts` property and replaced it with `EmbedFontsSubset` in Extended OCR engine
- `AsianOCREngine` is now obsolete and do not need to be set in Extended OCR engine

1.4 Bug fixes

1.4.1 OCR

- [SDK-112] Processing PDFs that have already been OCRed with another product in Native mode with `RemoveExistingPDFText` set to `True` causes double text layers when OCRed by Aquaforest OCR engines
- [SDK-122] - [Extended OCR] Running multi-threaded OCR through the SDK API throws System.AccessViolationException
- [SDK-124] - Standard OCR and Extended OCR engines cannot be used in the same project
- [SDK-125] - Very poor performance when processing certain PDF files
- [SDK-131] – Pages in certain PDF document go blank when processing with Standard OCR engine
- [SDK-135] - OCR SDK Hangs when processing in Native Mode
- [SDK-138] - [Extended OCR]: Ocr.DeleteTemporaryFiles() is not thread-safe
- [SDK-140] - Assembly load error when using both OCR engines in one application
- [SDK-141] - OCRing an image with Extended OCR and 'no OCR = true' causes "Invalid call to method" error
- [SDK-142] - Barcode is not detected if PerformPreprocessing is true or BlankPageThreshold > -1
- [SDK-150] - Visible text is rendered as unknown symbols when processing in PDF with custom encoding in Native mode
- [SDK-154] - Retrieving "Bits Per Component" from images in PDF pages throws exception

1.4.2 PDF Toolkit

- [PTK-1] - Provide Word to PDF feature
- [PTK-9] – License Pop Up issue on single page
- [PTK-10] - Trial version should limit text extraction to 3 pages
- [PTK-16] – Displays “Did not close PDF file” even if the document was not open in the first place
- [PTK-17] - Warning message 'You did not close a PDF Document' being displayed when processing secure PDF
- [PTK-25] - Not extracting the last item of text on a page
- [PTK-30] - Convert an image file to PDF: "alpha channel not implemented" error when processing JPEG File

1.5 New Sample Projects

1.5.1 OCR

- Sample project to demonstrate how to get text from PDFs with different searchability level. If the PDF is already searchable (i.e. contains text in all the pages), it will return the text without performing any OCR. If the PDF is not searchable, it will OCR it first and then get the text. This sample is available for all three OCR engines
- [SDK-127] Sample project to process color pages with Extended OCR engine
- [SDK-131] Sample project to demonstrate retrieving page confidence score with Standard OCR

1.5.2 PDF Toolkit

- Extract PDF Text as HOCR JSON
- Extract Text from Form
- Image Files merged to PDF
- Searchable PDF to Image PDF

1.5.3 Data Extractor

See Welcome page for Data Extractor samples.

1.5.4 Cloud OCR

See Welcome page for Cloud OCR samples.