

---

# Aquaforest Searchlight Release Notes



Version 1.10  
February 2016

# 1 Version 1.10

## 1.1 Enhancements

### 1.1.1 Updated Extended OCR engine

Aquaforest Searchlight 1.10 now has the latest version of the iDRS engine (iDRS 15) in the Extended OCR engine. It provides the following new features:

- Improved character recognition
- Additional output formats such as PDF/A-1a
- New Asian OCR engine
- JPEG2000 Compression

### 1.1.2 Re-image PDF

Both the Aquaforest and the Extended engines now have the option to re-image source PDF (also known as 'Convert to TIFF'), which rasterizes each page of the PDF document and add them to a new PDF with the OCR'd text layer.

### 1.1.3 Convert PDF to PDF/A

Previous versions of Aquaforest Searchlight only allowed converting TIFF files to PDF/A. With the newly added "Re-image PDF" option, PDF documents can also be converted to PDF/A.

### 1.1.4 Support for additional image types (BMP, JPEG and PNG)

This release of Aquaforest Searchlight can process BMP, JPEG and PNG files in addition to TIFF and PDF files.

The screenshot shows the Aquaforest Searchlight settings interface. The 'Document Settings' tab is active. The 'BMP Selection' section is highlighted with a red box and contains the following options:

- Process BMP Files: No
- Delete Original BMP: No

Other sections visible include:

- PDF Selection:** Process PDF Documents (Yes), Image Only PDFs (Yes), Partially Searchable (Yes), Fully Searchable (No), Hidden Text (Yes).
- TIFF Selection:** Process TIFF Files (No), Delete Original TIFF (No).
- JPEG Selection:** Process JPEG Files (No), Delete Original JPEG (No).
- PNG Selection:** Process PNG Files (No), Delete Original PNG (No).
- Filter Settings:** Temp Folder Location (C:\Aquaforest\Searchlight\temp), Filter Rule (No Filter), From (12/02/2015), To (12/02/2015), Exclude Specific Documents (checked).
- Document Error Settings:** Document Error Rule (Take no Action), Document Error Location.

### 1.1.5 Exclude specific documents

Specific documents can now be excluded from processing (both Audit and OCR). Documents to be excluded can be set through Filter Settings in the Document Settings page.

The screenshot shows the 'Document Settings' page in Aquaforest Searchlight. The 'Filter Settings' section is highlighted with a red box. It includes a 'Temp Folder Location' field with the value 'C:\Aquaforest\Searchlight\temp'. Below this, the 'Filter Rule' is set to 'No Filter'. The 'From' and 'To' date pickers are both set to '12/02/2015'. The 'Exclude Specific Documents' option is selected, indicated by a radio button. Other sections include 'PDF Selection', 'BMP Selection', 'JPEG Selection', and 'PNG Selection', each with 'Process' and 'Delete Original' options.

### 1.1.6 Temp Location

The temporary folder used to keep files before auditing and OCR can now be set through the UI rather than the Searchlight.config file.

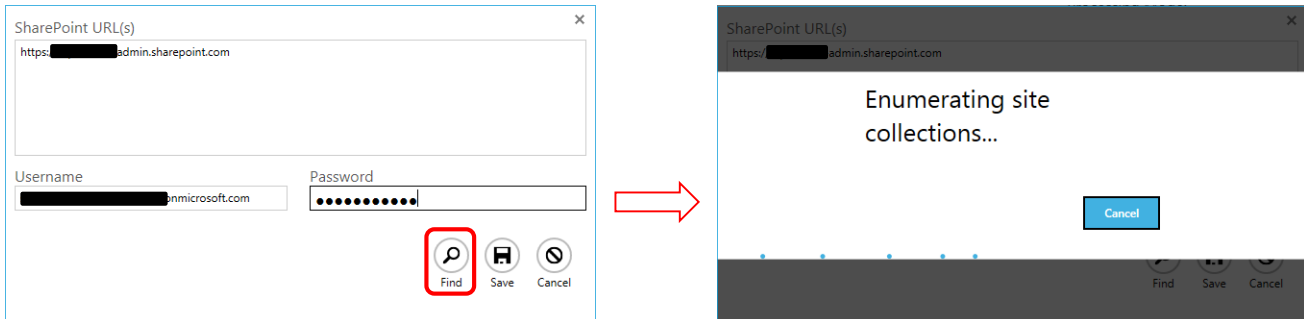
The screenshot shows the 'Document Settings' page in Aquaforest Searchlight. The 'Temp Folder Location' field is highlighted with a red box, showing the value 'C:\Aquaforest\Searchlight\temp'. The 'Filter Settings' section is also visible, with the 'Exclude Specific Documents' option selected. Other sections include 'PDF Selection', 'BMP Selection', 'JPEG Selection', and 'PNG Selection', each with 'Process' and 'Delete Original' options.

### 1.1.7 Active Directory Federation Service (AD FS) login

Aquaforest Searchlight now supports login to SharePoint Online (Office 365) configured to use AD FS.

### 1.1.8 Enumerate site collections

Aquaforest Searchlight can now enumerate site collections if the root admin URL is provided. This will facilitate adding multiple site collections at once. This feature is only available for Office 365.



### 1.1.9 Retrieve documents from SharePoint lists that exceed the List View Threshold

Aquaforest Searchlight can now get documents from SharePoint document libraries/lists that have more items than their List View Threshold.

### 1.1.10 Audit and OCR documents one by one

In previous versions of Aquaforest Searchlight, for SharePoint document libraries, all candidate documents were downloaded first before performing Audit and OCR. However, this required a considerable amount of free space in the local computer if the document library being processed was really big or if several document libraries were being processed at the same time.

In this release, documents are audited as soon as they are downloaded. If the processing mode is "Audit and OCR" and there is enough space in the local computer, the same downloaded documents can be used for OCR after all documents have been audited. However, if space is an issue, the documents can be deleted as soon as they have been audited and they will be downloaded again during the OCR process. To delete the documents after audit, the setting "deleteDocumentsAfterAudit" needs to be set to true in the Searchlight.config file.

### 1.1.11 Default OCR settings

In previous versions of Aquaforest Searchlight, OCR settings were hard-coded in the application. In this release, the OCR settings are loaded from the properties.xml file of the OCR engine being used.

- Aquaforest engine: "[installation path]\tj\bin\ocr\Properties.xml"
- IRIS (Extended) engine: "[installation path]\extendedocr\Properties.xml"

This can be useful if you have a set of OCR settings that work best for the type of documents you have and want to use the same OCR settings for all newly created document libraries.

**Note:** Aquaforest Searchlight does not modify the Properties.xml file. To set default values, you need to manually update the relevant Properties.xml file.

### 1.1.12 Ignore previously OCR'd documents

Searchlight may re-OCR documents that have already been processed previously if its modified date in SharePoint has changed since the last time it was processed and process "Fully Searchable" and/or "Partially Searchable" options are set in the Document Settings. The modified date can change if a document is replaced by a new one or its metadata/properties are modified in SharePoint.

To avoid re-processing these documents again irrespective of whether the modified has changed, set the "ignorePreviouslyOcredDocuments" setting to true in Searchlight.config. The default value is false.

### 1.1.13 Skip checked-out documents

It is now possible to skip checked-out documents from being processed (during OCR stage only). This is controlled by the "skipCheckedOutDocument" setting in Searchlight.config. The default value is true.

### 1.1.14 Retain Approval Status

When Aquaforest Searchlight processes documents in a SharePoint library which requires Content Approval, it will set them to 'Pending' after processing. To retain the original Approval Status after the documents have been processed, set the "retainApprovalStatus" setting to 'true' in Searchlight.config.

**Note:** If this setting is set to true, the "Retain Modified Date" in Aquaforest Searchlight will not work.

### 1.1.15 Audit Chart

A new feature has been added to allow users to view the audit results in a more user friendly graphical report as shown below. This report can be generated by going to Library → Status and click on the Report button.



# Audit Results

## SharePoint File Searchability

14 October 2015

Searchable  
Page Count

99%

### Locations:

http://Aquaforest001/Library1



*Run Searchlight with OCR enabled to make image or partially searchable PDFs fully searchable.*

*To see the full list of files that are not searchable go to the "run details" section of the Searchlight "Library" menu.*

<b>Document Library ID</b>	5
<b>Run ID</b>	30834
<b>Connection Totals</b>	Total Documents: 15 Total Error Documents: 0 Total Pages: 312 Total Searchable Pages: 309 (99 %)
<b>PDF Documents</b>	Total PDF Documents: 15 Image-only PDFs: 0 (0 %) Partially Searchable PDFs: 3 (20 %) Fully Searchable PDFs: 12 (80 %) Error PDF Documents: 0 Total PDF Pages: 312 Image-only Pages: 3 (1 %) Fully Searchable Pages: 309 (99 %)
<b>Image (TIFF,BMP,JPG,PNG) Documents</b>	Total Image Documents: 0 Error Image Documents: 0 Total Image Pages: 0

Aquaforest Searchlight 1.10.151014.0

Document Library ID: 5

Run ID: 30834

14 October 2015

### 1.1.16 Performance

The performance of several database heavy operations have been improved such as retrieving Run History/Details and deleting large document libraries.

### 1.1.17 Database Locks

When processing a document library using multiple cores, there used to be lots of "Database is locked" messages that were generated, which sometimes crashed the Aquaforest Searchlight service. This has been fixed in this release. However, it is still possible to get database locks when processing several document libraries at once using multicore but the frequency should be significantly reduced.

### 1.1.18 UI Changes

The following pages have been restructured to make them more user friendly:

- Library → OCR Settings
- Library → Run Details
- Library → Document Archive Settings

- Settings → License
- Settings → Theme

### 1.1.19 New Themes

There are now 23 different Accent colours to choose from both Light and Dark themes. The default is Light Blue.

The screenshot shows the Aquaforest Searchlight settings interface. At the top, there is a navigation bar with the logo and the text "AQUAFORST SEARCHLIGHT". Below this, there are navigation links: "Dashboard", "Library", "Settings" (highlighted in blue), and "Help & Support". Under "Settings", there are sub-links: "License", "Email", "Theme" (highlighted in blue), and "Advanced".

The "Theme" section is divided into two panels: "DARK THEMES" and "LIGHT THEMES". Each panel displays a grid of 23 color swatches with their corresponding names. In the "LIGHT THEMES" panel, the "Blue" swatch is highlighted with a blue background, indicating it is the selected theme.

DARK THEMES							
Red	Green	Blue	Purple	Orange	Lime	Emerald	Teal
Cyan	Cobalt	Indigo	Violet	Pink	Magenta	Crimson	Amber
Yellow	Brown	Olive	Steel	Mauve	Taupe	Sienna	


LIGHT THEMES							
Red	Green	Blue	Purple	Orange	Lime	Emerald	Teal
Cyan	Cobalt	Indigo	Violet	Pink	Magenta	Crimson	Amber
Yellow	Brown	Olive	Steel	Mauve	Taupe	Sienna	

## 2 Version 1.05

### 2.1 Enhancements

#### 2.1.1 Add Multiple SharePoint URLs

Multiple SharePoint URLs can now be added at once using the new enhanced Add New Location wizard. Each URL must be in a new line as shown below.



SharePoint URL(s)

http://mysharepoint/site1  
http://mysharepoint/site2  
http://mysharepoint/sites/sitecollection/site3

Username: username

Password: .....

Save Cancel

#### 2.1.2 Download Progress

The dashboard now displays the progress when downloading documents in the following format: "Downloading x of y".

#### 2.1.3 Download Retries

Occasionally, there might be some intermittent network problems which can cause problems when downloading files from SharePoint for processing. To cope with this, a retry mechanism has been implemented that will retry downloading in the event of such network problems. The number of retries and the amount of time to wait between retries can be controlled through the following config setting:

```
<add key="downloadRetries" value="5,1000" />
```

The value needs to be entered in the format "x,y", where x is the number of retries and y is the amount of time in milliseconds to wait for each retry.

This config setting can be found in the "Searchlight.config" file located at: "[installation path]\config\Searchlight.config".

#### 2.1.4 Database Update Retries

Sometimes, if a document library is set to process using multiple cores, Searchlight may encounter problems when it tries to update the database due to it being 'locked' because of concurrent updates. To overcome this problem, a retry mechanism has been implemented that will retry updating the database if it fails the first time. The number of retries and the amount of time to wait between retries can be controlled through the following config setting:



```
<add key="databaseRetries" value="5,1000" />
```

The value needs to be entered in the format "x,y", where x is the number of retries and y is the amount of time in milliseconds to wait for each retry.

This config setting can be found in the "Searchlight.config" file located at: "[installation path]\config\Searchlight.config".

### 2.1.5 Form-based authentication

Searchlight can now process SharePoint libraries that require form-based authentication.

### 2.1.6 Remove Hidden Text

Existing hidden text (text that was added as a result of a previous OCR) can now be removed from the PDF file so that the resulting searchable PDF file does not have two layers of the same text. This can be achieved by setting the "Remove Hidden Text" option to True.

### 2.1.7 Remove Visible Text

Visible text (text as a result of conversion from an electronic document such as Word to PDF) can now be excluded from the OCR process. This only affects engine 2 of Aquaforest OCR and the Extended OCR (IRIS engine).

To enable this feature:

- Aquaforest OCR - set "PdfToImageIncludeText" to False in properties.xml
- Extended OCR – set "Remove Visible Text" to True from General OCR Settings in the GUI.

### 2.1.8 Retain Creation/Modified Date/User

In this release of Aquaforest Searchlight, there is the extended functionality of retaining created date, modified user, created user and modified user of documents.

	Creation Date	Created User	Modified Date	Modified User
SharePoint	✓	✓	✓	✓
PDF metadata	✓	✓	✓	N/A
Windows File System	✓	✓	✓	N/A

"Create User" maps best to "Owner" in Windows File System metadata. For this to be manipulated Searchlight would need to be running with sufficient administrative privileges.

**Note:** Previous versions of Aquaforest Searchlight had two options "Retain SharePoint TIFF Creation Date" and "Retain Creation Date" which have now been merged to one option namely "Retain Creation Date". If any of the two options were set to 'True' in the previous version, it will be carried over to the new field.

### 2.1.9 Multicore support

In this version, the support for multicore processing has been increased from 8 cores to 64 cores.

## 2.2 Bug Fixes

### 2.2.1 SharePoint Template Types

In previous versions of Searchlight, only document libraries and lists with Server Template IDs 101 and 100 respectively were processed. As a result, document libraries and lists created using custom templates were skipped. This has now been fixed.