

# Aquaforest

## Searchlight Reference Guide



# Searchlight **Reference Guide**



Version 2.5  
January 2023

# Content

<b>1</b>	<b>PRODUCT OVERVIEW</b>	<b>5</b>
<b>1.1</b>	<b>The Business Problem: Documents that are not searchable.</b>	<b>5</b>
<b>1.2</b>	<b>The Solution: Aquaforest Searchlight</b>	<b>5</b>
<b>2</b>	<b>INSTALLATION AND LICENSING</b>	<b>6</b>
<b>2.1</b>	<b>System Requirements</b>	<b>6</b>
<b>2.2</b>	<b>SharePoint Online (Office 365) System Requirements</b>	<b>6</b>
<b>2.3</b>	<b>Licensing</b>	<b>6</b>
2.3.1	Entering License Keys	7
2.3.2	Service Configuration	7
<b>3</b>	<b>AQUAFOREST SEARCHLIGHT MODULES</b>	<b>9</b>
<b>3.1</b>	<b>Multi-core Module</b>	<b>9</b>
<b>3.2</b>	<b>OCR Engines Modules</b>	<b>9</b>
3.2.1	Standard OCR Module (Included with the standard product)	9
3.2.2	Extended (IRIS) OCR Module (Included with the standard product)	9
3.2.3	Extended OCR Asian Languages Module	9
3.2.4	Extended Arabic & Farsi Languages Module	9
3.2.5	Extended Hebrew Language Support	9
3.2.6	Extended OCR Advanced Compression	9
<b>4</b>	<b>SEARCHLIGHT ARCHITECTURE AND CONCEPTS</b>	<b>10</b>
<b>4.1</b>	<b>Supported Formats</b>	<b>11</b>
<b>4.2</b>	<b>Searchlight Libraries</b>	<b>11</b>
<b>4.3</b>	<b>Searchability Status</b>	<b>11</b>
<b>4.4</b>	<b>Audit and Candidate Identification</b>	<b>12</b>
<b>4.5</b>	<b>Document Stores Concepts</b>	<b>12</b>
4.5.1	SharePoint and Office 365 Document Stores Concepts	12
4.5.1.1	File and path lengths	12
4.5.1.2	Versioning	12
4.5.1.3	URL formats	13
4.5.2	Windows File System Stores Concepts	13
4.5.2.1	File and path lengths	13
4.5.2.1.1	Windows File System Standard Windows File System	13
4.5.2.1.2	Windows File System (Unicode)	13
4.5.2.1.3	Windows File System (long path)	14
4.5.2.2	File Access Permissions	14
4.5.3	Azure File Storage Stores Concepts	14
4.5.4	Azure Blob Storage Stores Concepts	14
4.5.5	Mixed Storage Types	14

<b>4.6</b>	<b>Archiving</b>	<b>15</b>
<b>4.7</b>	<b>Aquaforest Searchlight Service</b>	<b>15</b>
<b>5</b>	<b>QUICK START GUIDE</b>	<b>16</b>
<b>5.1</b>	<b>Creating a Library</b>	<b>16</b>
5.1.1	Library Settings	17
5.1.2	Document Settings	18
5.1.3	Document Archive Settings	20
5.1.4	OCR Settings	21
5.1.4.1	Extended OCR Engine Settings	21
5.1.4.2	Standard OCR Engine Settings	21
5.1.5	Scheduler	22
5.1.6	Alert Settings	22
5.1.7	Finish	24
<b>5.2</b>	<b>Updating a Library</b>	<b>25</b>
<b>5.3</b>	<b>Importing settings from an existing Library</b>	<b>26</b>
<b>5.4</b>	<b>Audit &amp; Conversion Status</b>	<b>27</b>
<b>6</b>	<b>THE AQUAFOREST SEARCHLIGHT TOOL</b>	<b>31</b>
<b>6.1</b>	<b>Welcome Screen</b>	<b>31</b>
<b>6.2</b>	<b>Dashboard</b>	<b>32</b>
<b>6.3</b>	<b>Library</b>	<b>33</b>
6.3.1	Library Status	33
6.3.2	Library Settings	33
6.3.3	Document Settings	35
6.3.3.1	Retain Creation/Modified Date/User	37
6.3.3.2	SharePoint Libraries	39
6.3.3.3	SharePoint Lists	40
6.3.4	Document Archive Settings	41
6.3.5	OCR Settings	42
6.3.5.1	Standard OCR Settings	42
6.3.5.1.1	General Settings	42
6.3.5.1.2	PDF Source Settings	43
6.3.5.1.3	Image Source Settings	45
6.3.5.2	Extended OCR Settings	46
6.3.5.2.1	General Settings	46
6.3.5.2.2	PDF Source Settings	47
6.3.5.2.3	Image Source Settings	48
6.3.5.2.4	Advanced Pre-processing Settings	50
6.3.6	Run Details	52
6.3.6.1	Run Details Context Menu	52
6.3.7	Scheduler Settings	53
6.3.8	Alert Settings	54
6.3.8.1	Action	54
6.3.8.2	Email	55
6.3.8.3	Report	56
6.3.8.4	Trigger	57
<b>6.4</b>	<b>Help &amp; Support</b>	<b>58</b>
6.4.1	Diagnostic Tool	59
6.4.2	Database Clean-up Tool	59

<b>6.5</b>	<b>Settings</b>	<b>61</b>
6.5.1	License Settings	61
6.5.2	Email Settings	62
6.5.2.1	SMTP	62
6.5.2.2	Azure OAuth2	63
6.5.3	Themes	64
6.5.4	Date & Time	64
6.5.5	Advanced Settings	64
<b>6.6</b>	<b>Searchlight.config file</b>	<b>65</b>
<b>7</b>	<b>ACKNOWLEDGEMENTS</b>	<b>69</b>

# 1 Product Overview

Aquaforest Searchlight is an in-place document processing tool that is designed to monitor and make files within an organization Searchable. It is able to integrate with Microsoft SharePoint and Windows File Systems.

## 1.1 The Business Problem: Documents that are not searchable.

Studies have shown that in most organizations over 20% of documents are not fully text searchable so will not be located by text search or discovery exercises. In addition, a greater percentage of documents may not be tagged with appropriate metadata. With the increase in distributed capture and ad-hoc publishing to document stores such as Microsoft SharePoint, there is a need for a solution to this problem that does not require a strict capture-time process.

Many types of documents are not searchable without special processing. For example:

- Scanned TIFF Files
- Image PDF Files
- Image Files (BMP, PNG, JPG)
- Faxes

These types of files need to be processed with Optical Character Recognition (OCR) technology to create a text version of the file contents which allows a searchable PDF to be created by merging the original page images with the text. The text is stored in the PDF file as a hidden layer overlaying each page image. This enables the file to be searched.

Documents stored in Microsoft SharePoint may often be lacking key metadata required to enable straightforward metadata searches. For example, attributes such as "Keywords" or "Company" may not have been fully indexed when the document was stored in SharePoint. The Aquaforest Searchlight Metadata Extractor module can be configured to automatically add metadata to new and existing documents.

In order to enable searches across files in SharePoint, Windows Search or other Document Management Systems the searchable files need to be indexed by the system. System iFilters manage this automatically for Microsoft Office but for PDF files a separate iFilter is required. A free iFilter is available from Adobe which does a good job but only indexes basic PDF content, not PDF titles, subjects, authors, keywords, annotations, bookmarks, attachments, create time/date, number of pages.

## 1.2 The Solution: Aquaforest Searchlight

- Audits document stores to determine which documents require processing
- Document Stores are monitored to deal with new and updated documents.
- Dashboard provides a convenient summary of the state of all managed stores
- Provides detailed conversion reporting.
- convenient GUI which enables management of all stores via a single interface
- OCR Support for 100+ languages including English, Spanish, German, French



## 2 Installation and Licensing

### 2.1 System Requirements

<b>Supported Operating Systems</b>	<ul style="list-style-type: none"><li>• Windows 10(x64)</li><li>• Windows Server 2012 R2 (x64)</li><li>• Windows Server 2016</li><li>• Windows Server 2019</li></ul>
<b>Supported Document Stores</b>	<ul style="list-style-type: none"><li>• SharePoint 2010</li><li>• SharePoint 2013</li><li>• SharePoint 2016</li><li>• SharePoint 2019</li><li>• SharePoint Online (Office 365)</li><li>• OneDrive for Business</li><li>• Azure File Storage</li><li>• Azure Blob Storage</li><li>• Windows File Systems</li></ul>
<b>Disk Space</b>	950 MB
<b>Memory</b>	Minimum 4GB (recommended 8GB)
<b>Visual C++ Redistributable</b>	Visual C++ 2017 Redistributable ( <a href="#">x86</a>   <a href="#">x64</a> )
<b>.NET Framework</b>	<a href="#">4.7.2</a>

### 2.2 SharePoint Online (Office 365) System Requirements

<b>Supported Operating Systems</b>	Windows 10 (x64) Windows Server 2012 (x64) Windows Server 2016 Windows Server 2019
<b>Additional tools</b>	SharePoint Server Client Components SDK ( <a href="#">x86</a>   <a href="#">x64</a> )

### 2.3 Licensing

Aquaforest Searchlight has 3 main licensing levels:

- Single Core
- 4 Cores
- 8 Cores

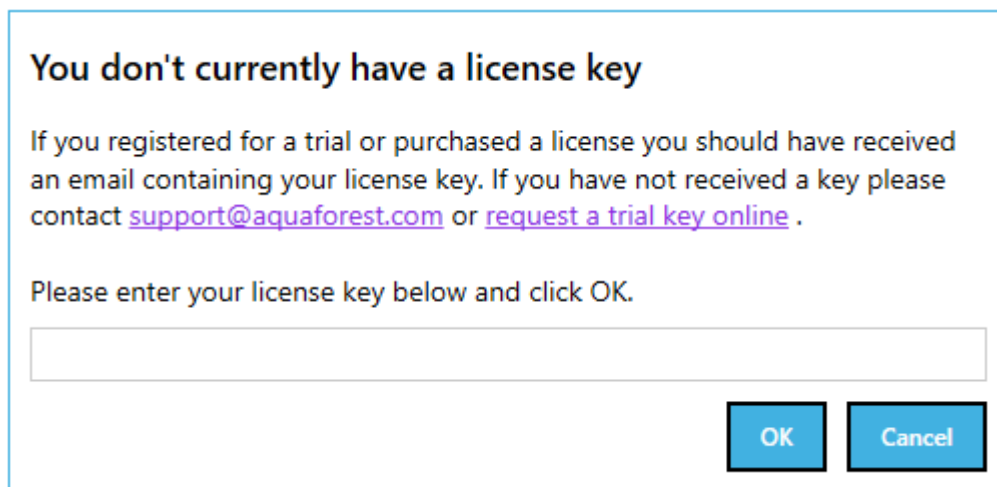
[Further Modules](#) are also available upon request. These are:

- Multi-core module with more than 8 cores. You can add additional blocks of 4 cores up to a maximum of 64
- Intelligent High-Quality Compression
- Asian Languages OCR support
- Arabic & Farsi Languages OCR support
- Hebrew Language OCR support

Trial licenses usually are time limited, that is, it will expire after a specified date or x days after installation. They may also limit the number of documents that can be OCR'd.

### 2.3.1 Entering License Keys

Aquaforest Searchlight will not run without a valid license key. If you do not have a valid license key, you will be prompted to enter a valid license key.



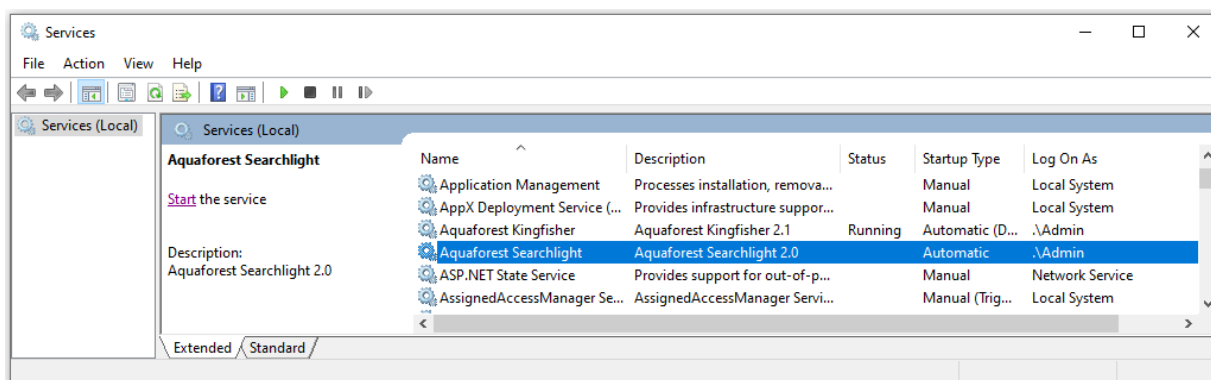
Email [support@aquaforest.com](mailto:support@aquaforest.com) to request a key if you do not have one. If you have a valid license key and wish to update it with a new one, go to **Settings > License** tab.

### 2.3.2 Service Configuration

The Aquaforest Searchlight Windows Service is required to log in with an account that has full administrative rights to the File System locations used for Aquaforest Searchlight File System libraries and File System locations used for [Errors](#), [Archives](#) and [Reports](#).

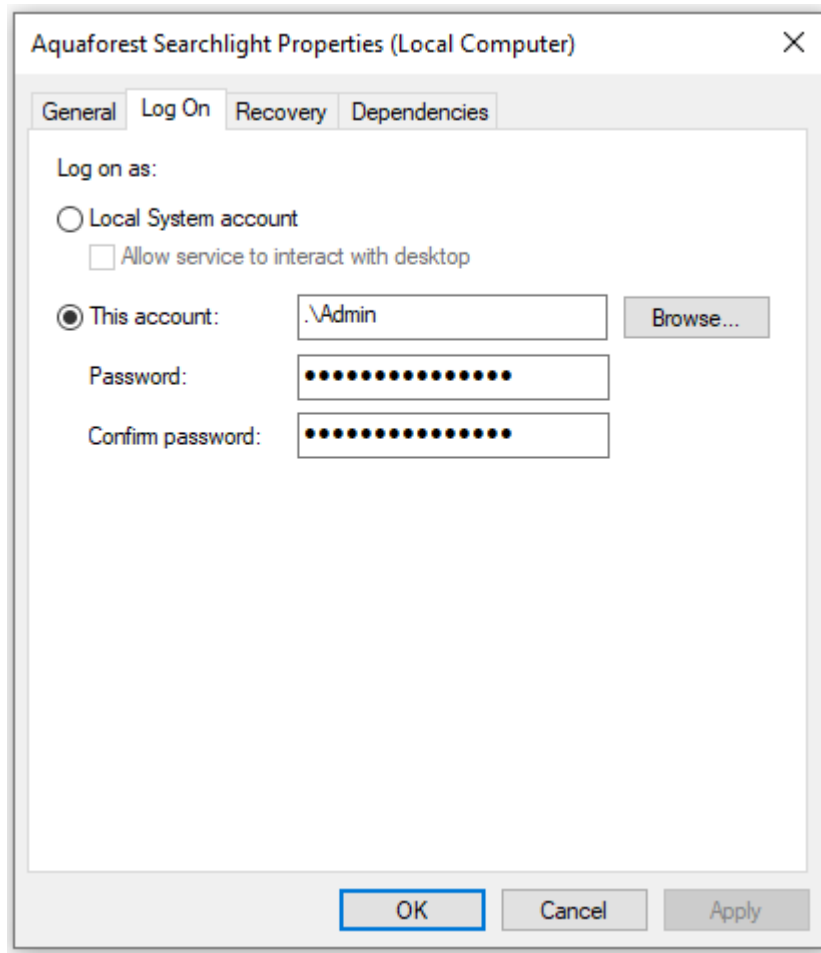
Configure Windows Service setting:

- Log-on to the computer as an Administrator.
- Either
  - From Control Panel, launch Administrative Tools.
  - From Administrative Tools, launch Services.
- Or
  - Search from the task bar for Services and launch Services:



- Select and double-click on the **Aquaforest Searchlight** service to bring up the **Aquaforest Searchlight Properties** dialog.
- Click the **Log On** tab. Select **This account** and type the username and password for the user for the service.





- Click OK to close the property dialog box and return to the main Services window. The service will not use the new user until it is started again.

Start (or Restart) the **Aquaforest Searchlight** Service.

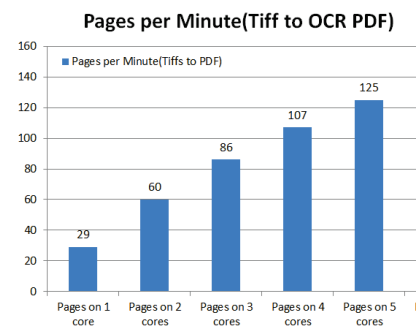
## 3 Aquaforest Searchlight Modules

### 3.1 Multi-core Module

This module is used to take full advantage of the number processors available on a computer.

The current release allows users to process up to 64 files in parallel.

The chart gives some indication of the improvement in throughput that can be expected when using the multi-core module.



### 3.2 OCR Engines Modules

OCR engines are the components that perform the task of text recognition on image files and extraction. Aquaforest Searchlight ships with two OCR Engines namely the Standard OCR Engine and the Extended (IRIS) OCR Engine. Below is an explanation of the OCR Engines.

#### 3.2.1 Standard OCR Module (Included with the standard product)

The Standard OCR Engine is included as a standard part of the product and can be used to convert Image PDFs and Images to searchable PDF documents. This engine has support of about 24 European Languages, but you can only OCR using one language at a time.



#### 3.2.2 Extended (IRIS) OCR Module (Included with the standard product)

The Extended Engine has the following benefits over and above the Standard OCR engine:

- Supports over 100 Languages.
- Support for multiple languages within a single document from the same alphabet e.g. French + German + Italian
- Canon IRIS OCR Engine - the same engine that is used in Adobe Acrobat
- Additional Advanced Pre-processing options for enhanced recognition, especially of poorer quality documents
- Optional Asian Languages Support
- Optional Arabic & Farsi Languages Support
- Optional Hebrew Language Support
- Optional iHQC Advanced PDF Compression



#### 3.2.3 Extended OCR Asian Languages Module

Adds support for Korean, Japanese, Simplified Chinese & Traditional Chinese languages.

#### 3.2.4 Extended Arabic & Farsi Languages Module

Adds support for Arabic and Farsi languages.

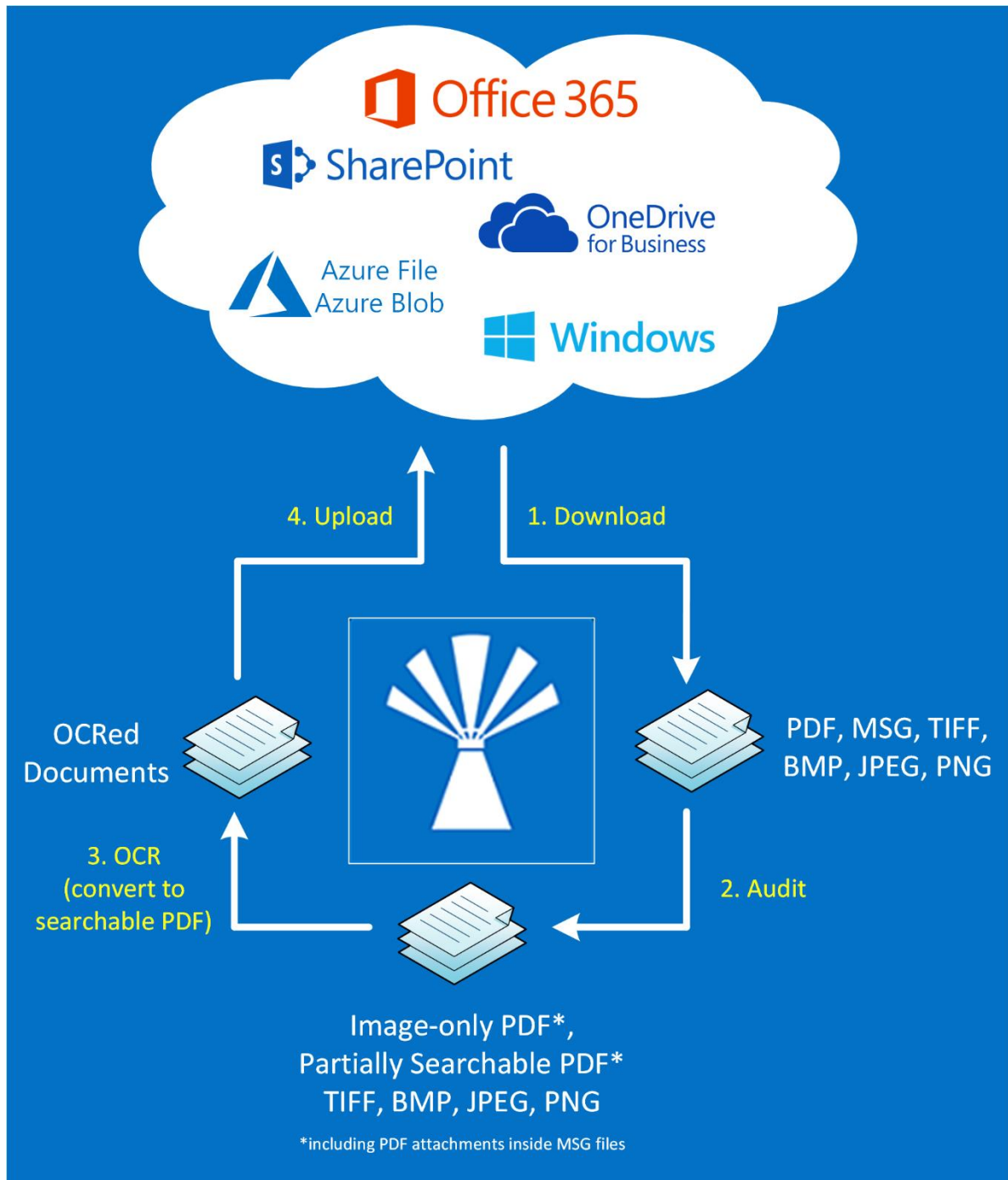
#### 3.2.5 Extended Hebrew Language Support

Adds support for Hebrew language.

#### 3.2.6 Extended OCR Advanced Compression

Aquaforest Searchlight uses IRIS's New Intelligent High-Quality Compression (IHQC). IHQC offers the most impressive PDF colour compression without compromising visual quality, text resolution and legibility of your documents. The IHQC module will be available if you purchase the IHQC license.

## 4 Searchlight Architecture and Concepts



There are 2 main stages when processing a Searchlight library, the Audit stage, and the OCR stage. At its most basic level, Aquaforest Searchlight will:

1. Audit Stage
  - 1.1. Download (SharePoint or Azure hosted locations) or copy (Windows file system locations) to a temporary local location.
  - 1.2. Analyse (Audit) the files to identify whether they need to be OCR'd.
  - 1.3. Record the results of the audit in the database.
2. OCR stage

- 2.1. If the file needs to be OCR'd then OCR it.
- 2.2. If the file has been OCR'd then replace the existing document (optionally restoring original file meta data and archiving the original)
- 2.3. Record the results of the OCR in the database.

Audits can be undertaken without the OCR stage to determine how many of your files are not currently searchable and allow you to determine the optimum way of fragmenting your libraries.

Audit (and OCR) results are recorded in a database which means that files which are unchanged do not need to be analyzed again, speeding up subsequent processing.

See the following [blog](#) for a more detailed explanation.

## 4.1 Supported Formats

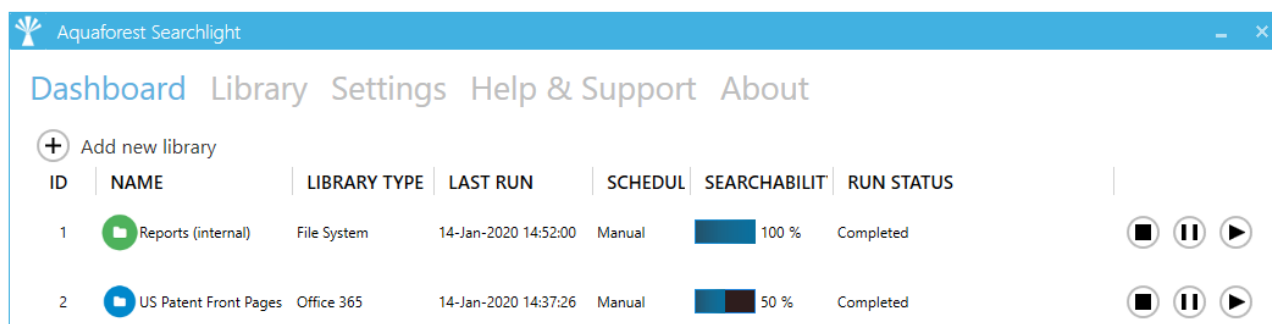
Aquaforest Searchlight currently supports TIFF, BMP, JPG, PNG and PDF documents (including PDF attachments inside MSG files) as input. As a result, candidate documents will always be one of these formats.

## 4.2 Searchlight Libraries

Aquaforest Searchlight revolves around the concepts of libraries. A Searchlight library can be described as a job in Aquaforest Searchlight that has all the settings required to process documents from specific Document Management Systems. It will usually consist of the following:

- The location(s) containing the documents that need to be processed.
- Document selection settings to indicate what types of documents to process (TIFF, PDF, etc.)
- OCR settings to use during the OCR phase

All Searchlight libraries are displayed in the Dashboard as shown below and the various settings associated with one can be accessed by double-clicking on it.



The screenshot shows the Aquaforest Searchlight Dashboard with a navigation bar containing 'Dashboard', 'Library', 'Settings', 'Help & Support', and 'About'. Below the navigation bar is a '+ Add new library' button. The main content area features a table with the following columns: ID, NAME, LIBRARY TYPE, LAST RUN, SCHEDULE, SEARCHABILITY, and RUN STATUS. There are two rows of data, each with a progress bar and control icons (stop, pause, play) on the right.

ID	NAME	LIBRARY TYPE	LAST RUN	SCHEDULE	SEARCHABILITY	RUN STATUS
1	Reports (internal)	File System	14-Jan-2020 14:52:00	Manual	100 %	Completed
2	US Patent Front Pages	Office 365	14-Jan-2020 14:37:26	Manual	50 %	Completed

A Searchlight library should not be confused with a SharePoint document library, which is a document library in SharePoint. See <https://www.aquaforest.com/wp/office-365-sharepoint-hierarchy-explained-2/> for a more detailed explanation.

## 4.3 Searchability Status

The searchability status of a document describes how indexable the document is. Searchlight will classify the searchability of documents in the following 3 categories:

### 1) Fully Searchable

A PDF document is fully searchable if all its pages have text that can be indexed and searched.

## 2) Partially Searchable

A partially searchable document contains some pages with text, others with only images or no images and no text (blank)

## 3) Image-only

This is a PDF that has been created from one or more images – most commonly because of scanning a document either directly to PDF or by converting a scanned TIFF image to PDF. These files do not contain any searchable text and most often comprise a set of Group4 or JBIG2 images in a PDF “wrapper”.

Image documents (TIFF, BMP, JPG and PNG) are always identified as image-only.

## 4.4 Audit and Candidate Identification

Before processing a document library, Aquaforest Searchlight will perform an Audit (analysis) on the document library to determine which documents are candidates for processing by examining each document’s searchability status and comparing it with the document selection settings in the **Library > Document Settings** tab.

## 4.5 Document Stores Concepts

### 4.5.1 SharePoint and Office 365 Document Stores Concepts

Aquaforest Searchlight can be configured to monitor multiple SharePoint libraries. Below are some concepts that should be taken into consideration during configuration.

#### 4.5.1.1 File and path lengths

The file path is everything after the server’s name and port number in the URL. File path includes the name of the site and subsites, document library, folders, and the file name itself.

SharePoint Type	Maximum file path Length	Maximum file or folder name length
SharePoint Online (Office 365)	400	400
SharePoint On-Premises 2019	400	400
SharePoint On-Premises 2016	256	128
SharePoint On-Premises 2013	256	128
SharePoint On-Premises 2010	256	128

#### 4.5.1.2 Versioning

Since Aquaforest Searchlight uses in-place processing, the source document is replaced by the resulting PDF file. However, if versioning is turned on, the resulting PDF file will be created as another version of the input file in SharePoint. If versioning is turned off, then the resulting PDF file replaces the source file.

### 4.5.1.3 URL formats

Below are examples of SharePoint URL formats accepted by Searchlight when setting up a document library. NOTE: Make sure the URLs start with "http" or "https"

#### Example formats

*Site/Web:*

- <https://myCompany>
- <https://myCompany/sites/mySite>
- <https://myCompany/sites/mySite/mySubSite>

*Document Library:*

- <https://myCompany/myLibrary>
- <https://myCompany/sites/mySite/myLibrary>
- <https://myCompany/sites/mySite/mySubSite/myLibrary>

*List:*

- <https://myCompany/Lists/myList>
- <https://myCompany/sites/mySite/Lists/myList>

*OneDrive for Business*

- [https://myCompany-my.sharepoint.com/personal/firstname\\_lastname\\_mycompany\\_onmicrosoft\\_com](https://myCompany-my.sharepoint.com/personal/firstname_lastname_mycompany_onmicrosoft_com)
- [https://myCompany-my.sharepoint.com/personal/firstname\\_lastname\\_mycompany\\_onmicrosoft\\_com/myLibrary](https://myCompany-my.sharepoint.com/personal/firstname_lastname_mycompany_onmicrosoft_com/myLibrary)

However, if the full URL is entered (i.e., ending with ".aspx") as shown below, Searchlight will try to automatically format it to one of the above accepted formats:

- <https://myCompany/sites/mySite/SitePages/Home.aspx>
- <https://myCompany/sites/mySite/myLibrary/Forms/AllItems.aspx>
- [https://myCompany/sites/mySite/\\_layouts/15/start.aspx#/myLibrary/Forms/AllItems.aspx](https://myCompany/sites/mySite/_layouts/15/start.aspx#/myLibrary/Forms/AllItems.aspx)
- <https://myCompany/sites/mySite/Lists/myList/AllItems.aspx>
- [https://myCompany/sites/mySite/\\_layouts/15/start.aspx#/Lists/myList/AllItems.aspx](https://myCompany/sites/mySite/_layouts/15/start.aspx#/Lists/myList/AllItems.aspx)
- [https://myCompany-my.sharepoint.com/personal/firstname\\_lastname\\_mycompany\\_onmicrosoft\\_com/\\_layouts/15/onedrive.aspx](https://myCompany-my.sharepoint.com/personal/firstname_lastname_mycompany_onmicrosoft_com/_layouts/15/onedrive.aspx)
- [https://myCompany-my.sharepoint.com/personal/firstname\\_lastname\\_mycompany\\_onmicrosoft\\_com/myLibrary/Forms/AllItems.aspx](https://myCompany-my.sharepoint.com/personal/firstname_lastname_mycompany_onmicrosoft_com/myLibrary/Forms/AllItems.aspx)

## 4.5.2 Windows File System Stores Concepts

### 4.5.2.1 File and path lengths

#### 4.5.2.1.1 Windows File System Standard Windows File System

The maximum length of a path is 260 characters (D:\some 256-character path string<NUL>).

#### 4.5.2.1.2 Windows File System (Unicode)

The Windows API has many functions that also have Unicode versions to permit an extended-length path for a maximum total path length of 32,767 characters.

This type of path is composed of components separated by backslashes, each up to 255 characters.

To specify an extended-length path, use the "\\?\\" prefix. For example, "\\?\D:\very long path".

#### 4.5.2.1.3 Windows File System (long path)

Starting in Windows 10 version 1607 it is possible to opt out of the MAX\_PATH limitations in common Win32 file and directory functions.

#### 4.5.2.2 File Access Permissions

The Aquaforest Searchlight Service must be configured with the security credentials of a user that has permissions to access that specific location.

#### 4.5.3 Azure File Storage Stores Concepts

The entire path, including the file name, must contain fewer than 2,048 characters.

The path is composed of components separated by backslashes (for example \\A\\B\\C\\D, each letter is a component), each component can be up to 255 characters in length.

#### 4.5.4 Azure Blob Storage Stores Concepts

Blob storage is a flat storage scheme. Within one container, each blob name identifies a blob. It is possible to simulate a folder structure using delimiters within the blob name.

Blobs are identified by both a container name and a blob name.

Container names are between 3 and 63 characters in length.

A blob name must be at least one character long and cannot be more than 1,024 characters long.

#### 4.5.5 Mixed Storage Types

Though it is possible within a Searchlight library to use one document store type as the source, and another document store type for both [Archive](#) location, and for [files generating errors](#), there will be issues due to differences in file path lengths and characters acceptable in file paths.

	Archive					Error				
	Windows File System	SharePoint Online (Office 365)	SharePoint On-Premises 20nn	Azure Blob Storage	Azure File Share	Windows File System	SharePoint Online (Office 365)	SharePoint On-Premises 20nn	Azure Blob Storage	Azure File Share
Source										
Windows File System	Green	Grey	Grey	Grey	Grey	Green	Grey	Grey	Grey	Grey
SharePoint Online (Office 365)	Yellow	Green	Grey	Grey	Grey	Yellow	Green	Grey	Grey	Grey
SharePoint On-Premises			Green	Grey	Grey			Green	Grey	Grey
Azure Blob Storage				Green	Grey				Green	Grey
Azure File Share					Green					Green

For general use, it is recommended that a Searchlight Library uses the same type of storage for all locations.

Use of Windows File System for Archive and Error locations has been tested, but there are issues with respect to path lengths and accepted characters as noted [above](#).

## 4.6 Archiving

To avoid making inadvertent changes to the source document, it is recommended to turn Archiving on to maintain a backup of the source documents.

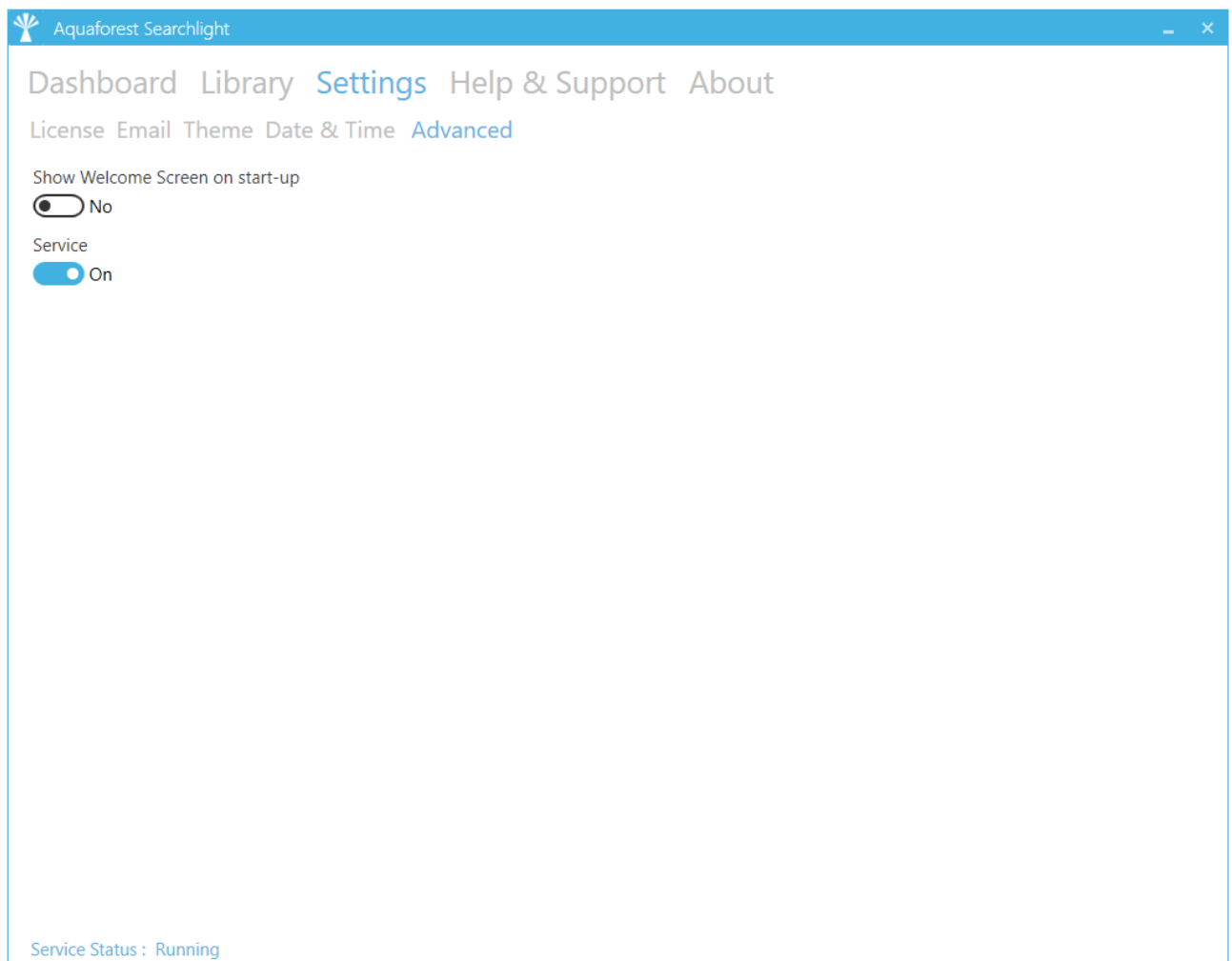
If Archiving is turned on, a copy of the file is created in a user specified archive location before any processing takes place. There is an option to retain the folder structure in the archive location.

## 4.7 Aquaforest Searchlight Service

This is the heart of the product and controls the execution of all libraries. Without it running, a library cannot be audited or OCRed. It is also used by the scheduler to automate the processing of libraries at regular time intervals without interfering with other work being performed on the machine it is installed in. It is also used to generate scheduled reports and sending email alerts.

The service can be turned on or off by going to **Settings > Advanced** tab.

The Service Status is displayed at the bottom left of all tabs.

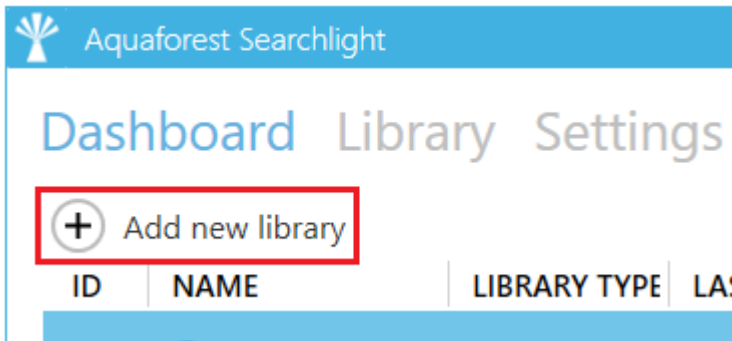




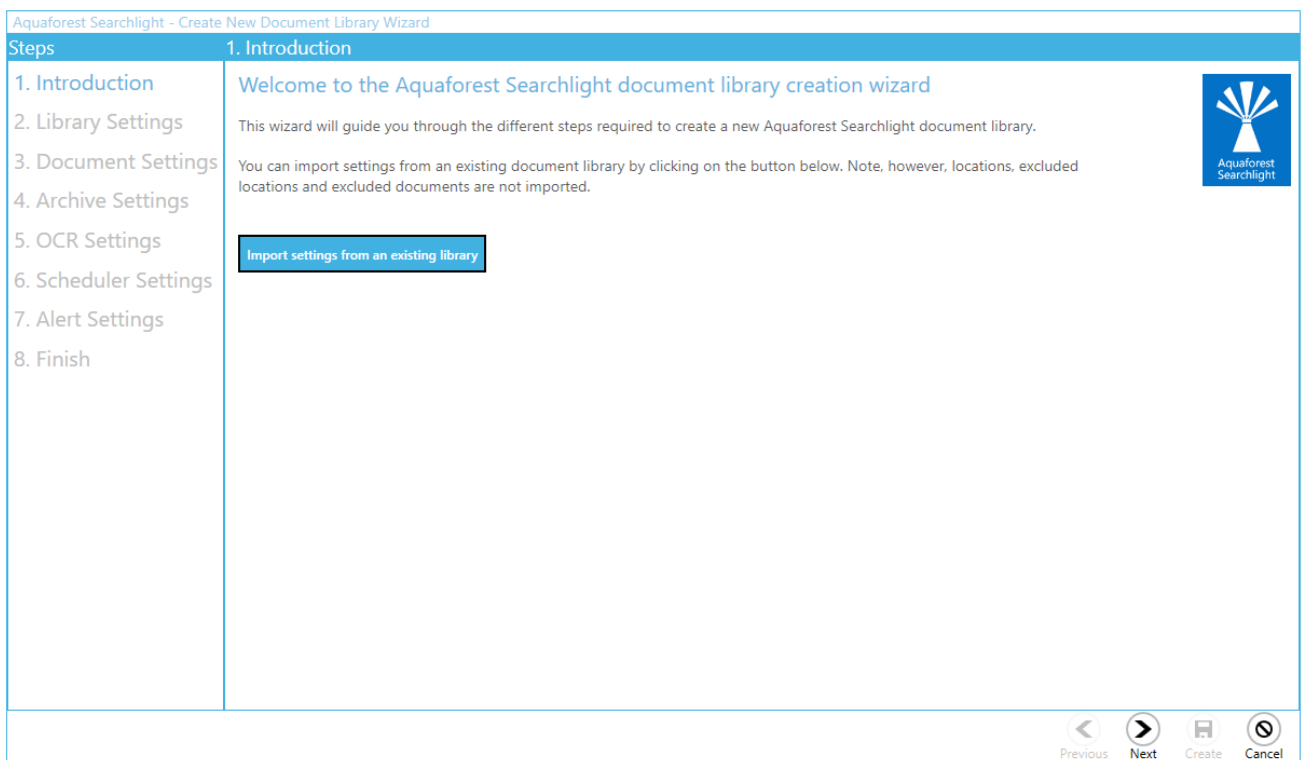
## 5 Quick Start Guide

### 5.1 Creating a Library

Creating Document Libraries in Aquaforest Searchlight is managed by a wizard. This wizard can be launched by clicking the **Add new library** button on the Dashboard.



The wizard provides helpful information throughout the different stages of the document library creation process which aids in better understanding the various steps and settings involved. Refer to [section 6.3](#) for detailed description of each of the settings in each page.



## 5.1.1 Library Settings

Aquaforest Searchlight - Create New Document Library Wizard

Steps

2. Library Settings

1. Introduction

2. Library Settings

3. Document Settings

4. Archive Settings

5. OCR Settings

6. Scheduler Settings

7. Alert Settings

8. Finish

Library Name: \_\_\_\_\_

Library Type: SharePoint On-Premises (1)

Locations: \_\_\_\_\_ (2) + Add new Location

Choose Library Icon: \_\_\_\_\_

Processing Mode: (4)

Audit Only

Audit and OCR

Cores: (5)

1

SharePoint Settings

Process SharePoint Lists: (6)

No

If versioning is off: (7)

Turn versioning on

Publish Major Version: (7)

Yes

Check-in Comment: (8)

OCR'ed by Aquaforest Searchlight on %DATE% %TIME%

Custom Check-in Column: \_\_\_\_\_ Comment: \_\_\_\_\_

Exclude Specific Locations


\*Filter Locations by Regular Expression (3)


Previous Next Create Cancel


- 1) Select the document source from the following: File System; SharePoint on-Premises; SharePoint Online (Office 365); Azure Blob Storage; Azure File Storage
- 2) Add new location(s) (depending on library type)
  - SharePoint On-Premises and SharePoint Online (Office365) locations can include one or more from:
    - SharePoint site collections
    - SharePoint sites
    - SharePoint document libraries
    - SharePoint lists.
  - one or more File System paths
  - one or more Azure Blob Storage paths
  - one or more Azure File Storage paths
- 3) There are 2 ways to filter locations:
  - a) Excluding specific locations – locations that match the specified site or library URL(s) are excluded.
  - b) By regular expressions – locations (site and library URLs) that match the specified regular expressions are included.

This is useful if you are processing a whole site collection and want to excluded specific locations and/or include only specific sites or libraries. For instance, you may want to only process sites and libraries containing the word “Resources” in their URL:

Only process locations whose URL match any of the following conditions:



 Add new condition

 OK

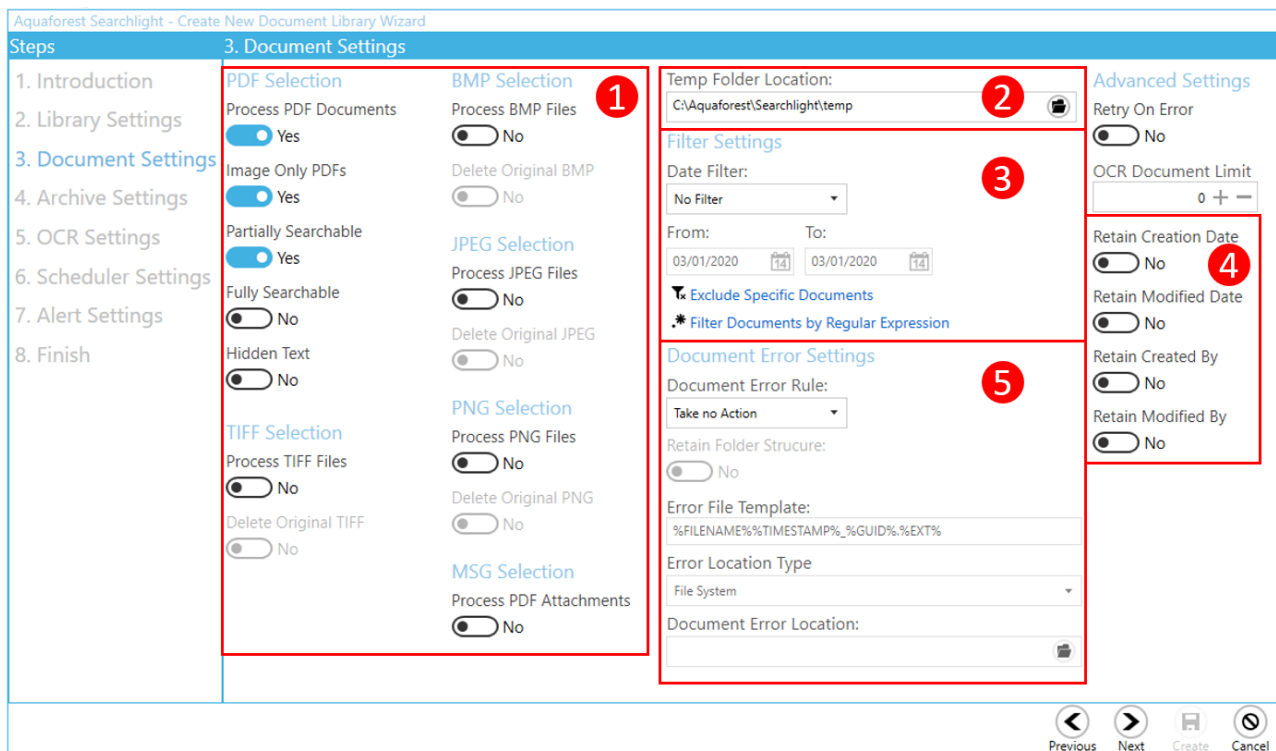
Below are a few useful resources to get you started with regular expressions:

- <https://docs.microsoft.com/en-us/dotnet/standard/base-types/regular-expression-language-quick-reference>
- <https://msdn.microsoft.com/en-us/library/ms972966.aspx>
- <https://msdn.microsoft.com/en-us/library/ff650303.aspx>

- 4) Do you only want to **Audit Only**, or **Audit and OCR**? Audit means that Searchlight will analyse the searchability of the documents and report how many searchable, partially searchable, and image-only documents are found in the location(s) specified, while Audit and OCR will find the non-searchable documents, and then make them searchable.
- 5) The number of cores to use to process documents in parallel. For instance, if 8 cores are specified, Searchlight will process 8 documents simultaneously, which will significantly reduce the total processing time. The hardware and license will have to support multiple cores.
- 6) Choose whether to process SharePoint Lists or not. If this is turned on, Searchlight will process the attachments in each list item. Note, however, that processing SharePoint lists can be extremely time consuming if they are very large.
- 7) Turn versioning on if you want to have a 'backup' of the original documents, otherwise the documents will be overwritten with new searchable ones (see also the [Archive Settings](#) step).
- 8) You can choose to add a check-in comment to the OCR'd files once they are uploaded to SharePoint. Optionally, you can also add a custom comment to a custom SharePoint column. However, the custom SharePoint column must be either of 'Text' or 'Date' type.

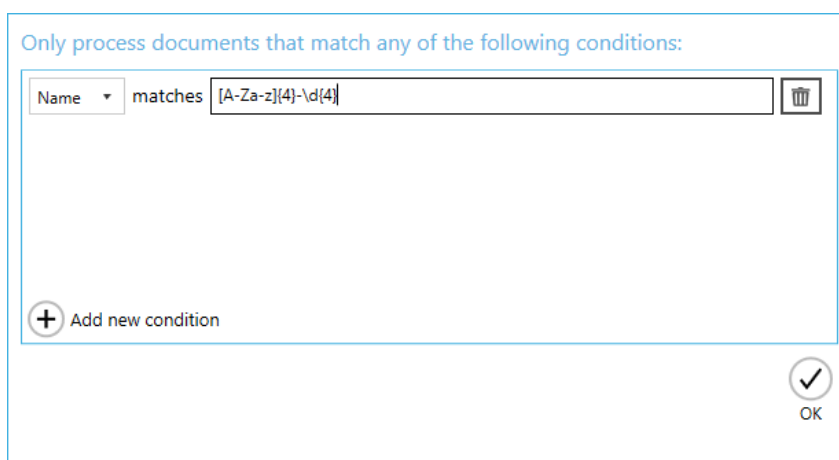
### 5.1.2 Document Settings

This page enables the user to specify rules and criteria for the selection of documents to be processed.



1. Select the document types to process. For image files, there is an option to delete the original images from the source location after they have been converted to searchable PDFs.
2. The **Temp Folder Location** is where Searchlight temporarily stores downloaded files as well as files created during OCR.
3. There are different options to filter documents:
  - a. By modified or creation date – documents that fall within the specified range are excluded.
  - b. By document paths – documents that match the specified paths are excluded.
  - c. By regular expressions - documents whose properties match the specified regular expressions are included

For instance, you may want to only process documents with the name format “ABCD-1234”:



4. There is also the option of retaining the original metadata on the document and in SharePoint so that even after uploading the searchable PDF these columns will not be changed.
5. If there is an error while processing a document, there are options to copy or move the file to an Error location. The folder structure of the source file can be retained.

### 5.1.3 Document Archive Settings

This page provides the option of archiving source files before OCR is applied to them, so there is a backup. The source folder structure can be retained in the archive folder.

Aquaforest Searchlight - Create New Document Library Wizard

Steps

4. Archive Settings

Document Archive Settings

Archive source images to Archive Folder:  Yes 1

Archive source PDF & MSG files to Archive Folder:  Yes

Retain folder structure:  No 2

Archive Rule:  3

Archive Template: %FILENAME%%TIMESTAMP%\_%GUID%.%EXT%

Location Type:  4

Archive Location: C:\sl2test\Archive

Previous Next Create Cancel

1. Select whether you want to archive just image files (TIFF, BMP, JPG and PNG) or PDF and MSG files.
2. Select if you want to retain the existing folder structure within the archive.
3. Select the archive rule – Copy to Archive Folder
4. Select the archive filename format, storage type (File System, SharePoint On-Premises, SharePoint Online, Azure Blob Storage, Azure File Storage) and location.

## 5.1.4 OCR Settings

In this section, you can set the OCR settings. Aquaforest Searchlight comes bundled with two OCR Engines: [Standard OCR engine](#) and the [Extended IRIS \(Canon\) OCR engine](#). The Extended OCR is the default engine and supports more languages (120+) than the Standard OCR engine. The Extended OCR engine can also process documents that have pages in different languages. See [section 3.2](#) for more information about the OCR engines.

### 5.1.4.1 Extended OCR Engine Settings

Aquaforest Searchlight - Create New Document Library Wizard

**Steps**

1. Introduction
2. Library Settings
3. Document Settings
4. Archive Settings
5. OCR Settings
6. Scheduler Settings
7. Alert Settings
8. Finish

### 5. OCR Settings

OCR Engine:  
 Aquaforest  Extended (IRIS)

General Settings PDF Source Settings Image Source Settings Advanced Preprocessing Settings

Auto Rotate  Off Despeckle 4 Remove Blank Pages  1 + -

Deskew  No Advanced Despeckle No Despeckle Interpolate  No

Remove Dark Borders  No Remove White Pixels  No Interpolation Mode Fast

Keep Original Image  Yes Work Depth 0 + - Interpolation Value 300 + -

Advanced Flags

Select Language(s)

- Dutch
- English
- Esperanto
- Estonian
- Faroese
- Fijian
- Finnish
- French
- Frisian
- Friulian

English, French, German

Previous Next Create Cancel

### 5.1.4.2 Standard OCR Engine Settings

Aquaforest Searchlight - Create New Document Library Wizard

**Steps**

1. Introduction
2. Library Settings
3. Document Settings
4. Archive Settings
5. OCR Settings
6. Scheduler Settings
7. Alert Settings
8. Finish

### 5. OCR Settings

OCR Engine:  
 Aquaforest  Extended (IRIS)

General Settings PDF Source Settings Image Source Settings

Auto Rotate  Off Despeckle 3

Deskew  On OCR Language English

Remove Lines  Off Box Graphics Treat all Graphics Areas as Text

Stamps

Advance Flags

Previous Next Create Cancel

## 5.1.5 Scheduler

The scheduler allows Aquaforest Searchlight to automate the running of document libraries. You can either run it manually, or run periodically, every day at a specified time or every hour etc.

The screenshot shows the '6. Scheduler Settings' step of the 'Create New Document Library Wizard'. On the left, a 'Steps' sidebar lists: 1. Introduction, 2. Library Settings, 3. Document Settings, 4. Archive Settings, 5. OCR Settings, 6. Scheduler Settings (highlighted), 7. Alert Settings, and 8. Finish. The main content area has three radio button options: 'Manual' (selected), 'Once per day', and 'Continuous'. Under 'Manual', there is an 'At:' field with a time picker set to 09:45:00. Under 'Continuous', there is an 'Every:' field with a numeric input '1', a '+' and '-' button, and a 'Hours' dropdown menu. Below that are 'Between' and 'And' time range pickers, with 'Between' set to 00:01:00 and 'And' set to 23:59:00. Under 'Run once', there is an 'On:' date picker set to 14/10/2016 and an 'At:' time picker set to 09:45:00. At the bottom right, there are four navigation buttons: 'Previous', 'Next', 'Create', and 'Cancel'.

## 5.1.6 Alert Settings

The alert settings provide you with the option of periodically sending email alerts as well as generating reports of job runs within a specified date range. Creating alerts is managed by another wizard within the library creation wizard.

1. Select the action(s) you want to perform.

The screenshot shows the '7. Alert Settings' step of the 'Create New Document Library Wizard'. The 'Steps' sidebar on the left highlights '7. Alert Settings'. The main content area is split into two columns: 'Configuration' and 'Action'. Under 'Configuration', there are three sections: 'Action' with 'Email' and 'Report' options, 'Trigger' with a 'Finish' option, and 'Finish'. Under 'Action', there is a question: 'What action(s) you want the alert task to perform?'. Below this are four toggle switches: 'Send an email' (Yes), 'Generate a CSV report' (Yes), 'Attach the CSV report to the email' (Yes), and 'Save Report' (No). At the bottom of the 'Action' column is a 'Location:' text input field with a file icon. At the bottom right, there are four navigation buttons: 'Previous', 'Next', 'Create', and 'Cancel'.

2. Select the email settings.

The screenshot shows the 'Alert Settings' step of the 'Create New Document Library Wizard'. The left sidebar lists steps 1 through 8, with '7. Alert Settings' selected. The main area is titled '7. Alert Settings' and has a 'Configuration' tab. Under 'Configuration', the 'Action' is set to 'Email'. The 'Action > Email' section contains the following fields: 'From Email Address' (support@aquaforest.com), 'To Email Address' (support@aquaforest.com), 'Email Subject' (%LIBRARYNAME% %STATUS%), and 'Email Message' (Further processing of '%LIBRARYNAME%' has been suspended. Log file: %LOGFILEPATH%). There is a 'Test Email' checkbox which is checked. Navigation buttons for 'Previous', 'Next', 'Create', and 'Cancel' are visible at the bottom.

3. Select the report settings. You can choose to get a summary of the library status as a whole and/or details about specific runs.

The screenshot shows the 'Alert Settings' step of the 'Create New Document Library Wizard'. The left sidebar lists steps 1 through 8, with '7. Alert Settings' selected. The main area is titled '7. Alert Settings' and has a 'Configuration' tab. Under 'Configuration', the 'Action' is set to 'Report'. The 'Action > Report' section contains the following settings: 'Library Audit Summary' (The library audit summary will contain statistics about current searchability status of the library as a whole as well as individual statistics about each document type in the library. Show library audit summary in report: Yes), 'Run Details Summary (OCR only)' (The run details will contain a summary of all the documents that were processed in a particular run: - No. of documents OCR'd, - No. of documents that failed to OCR, - etc... Show run details summary in report: Yes), and 'Show details of individual documents that were processed' (No). There is a note to 'Choose the columns that will appear in the report.' Navigation buttons for 'Previous', 'Next', 'Create', and 'Cancel' are visible at the bottom.

4. Select when you want the task to run. Based on the current settings, you will get an email with the report attached sent to the recipient every last Friday of the month at 8 am.



Aquaforest Searchlight - Create New Document Library Wizard

**Steps**

1. Introduction
2. Library Settings
3. Document Settings
4. Archive Settings
5. OCR Settings
6. Scheduler Settings
7. Alert Settings
8. Finish

**7. Alert Settings**

**Configuration**

Action

Email

Report

Trigger

Finish

**Trigger**

When do you want the task to start?

At 08:00, on the last Friday of the month.

Start: 17/10/2016 08:00:00 (14)

Month(s): January, February, March, April, May, June, Ju

Day(s):

Daily

Weekly

Monthly

One time

The: Last Friday

**Advanced Settings**

On Job Success

No

On Job Error

No

Expires

← →

← → ⏏ ⊗

### 5.1.7 Finish

On the **Finish** page, you will get a summary of all the settings you selected for this library. You can review them to see if you missed anything. If not, click on the **Create** button at the bottom of the wizard to create the library.

Aquaforest Searchlight - Create New Document Library Wizard

**Steps**

1. Introduction
2. Library Settings
3. Document Settings
4. Archive Settings
5. OCR Settings
6. Scheduler Settings
7. Alert Settings
8. Finish

**8. Finish**

**Summary**

**Library Settings**  
 Document Library Name: Test Library  
 Document Library Type: SharePoint  
 Location(s):  
 - http://Aquaforest001/Library1  
 Processing Mode: Audit and OCR  
 Audit History: 5  
 Cores: 8

**SharePoint Settings**  
 Versioning: Turn versioning on  
 Check-In Comment: OCR'ed by Aquaforest Searchlight

**Document Settings**

**PDF Documents**  
 Process PDF Documents: Yes  
 Image Only: Yes  
 Fully Searchable: No  
 Partially Searchable: Yes  
 Hidden Text: Yes

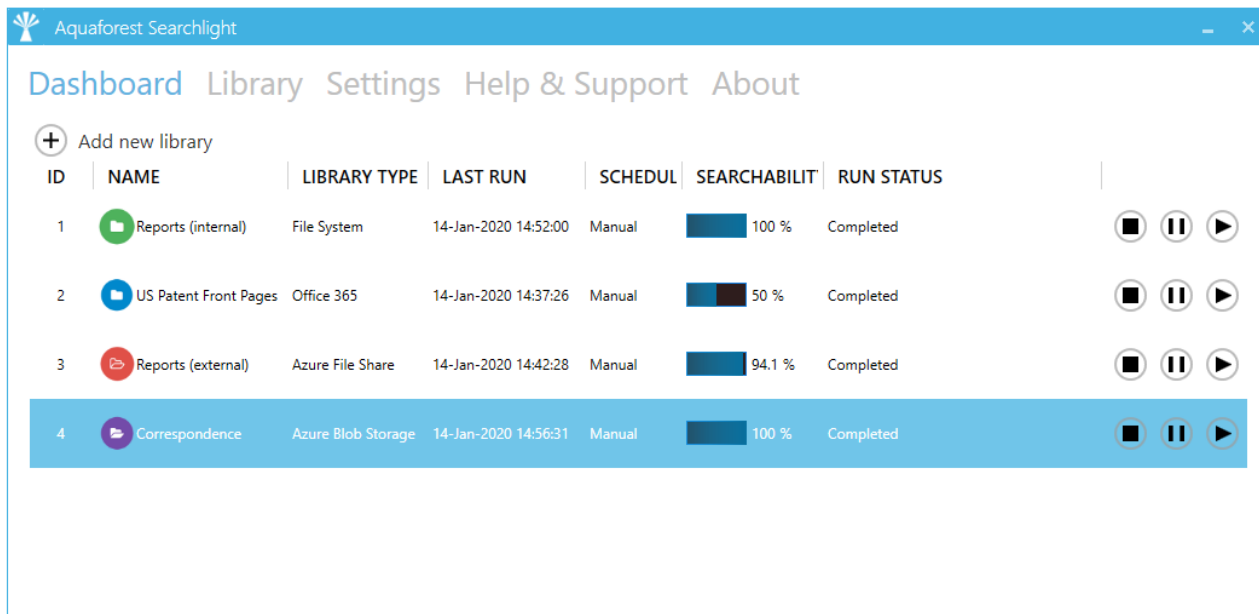
**TIFF Documents**  
 Process TIFF Documents: No  
 Delete Original TIFF Documents: No

**BMP Documents**  
 Process BMP Documents: No  
 Delete Original BMP Documents: No

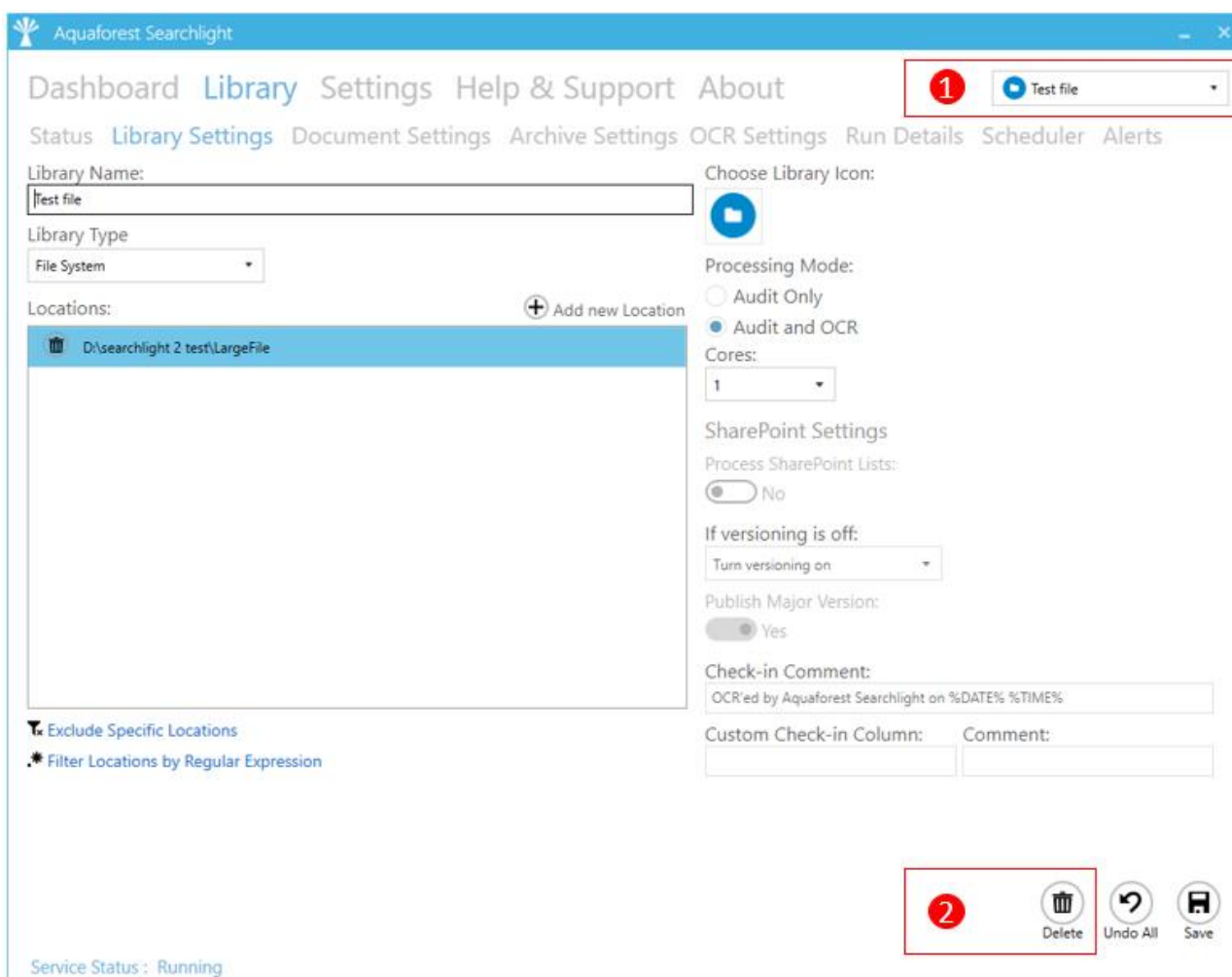
**IDOC Documents**

← → ⏏ ⊗

The new library will be added to the dashboard. As the library is set to run manually, click on the **Run** button to start processing.



## 5.2 Updating a Library

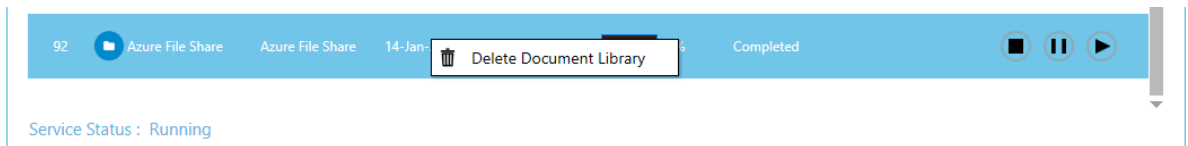


All the settings of a library can be edited by double-clicking the library from the dashboard, or by selecting the library and clicking on the Library Tab.

1. You can also select a library to edit by choosing the library from the combo box at the top of the page.

2. To delete the library, click on the **Delete** button at the bottom of the **Library Settings** page.

You can also, delete the library by right-clicking on the library from the dashboard and clicking on **Delete Document Library**

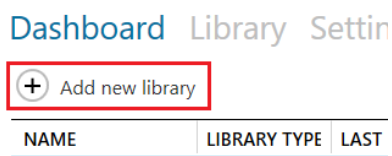


### 5.3 Importing settings from an existing Library

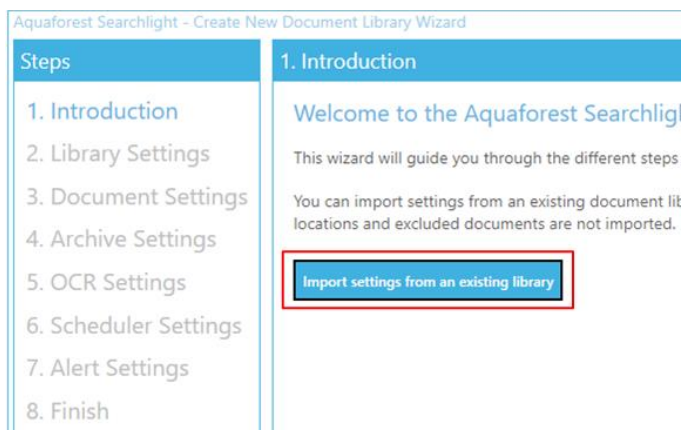
Searchlight also provides the ability to import settings from an existing library. However, “locations”, “excluded locations” and “excluded documents” are not imported because it is not allowed to have the same locations in multiple libraries.

To import settings:

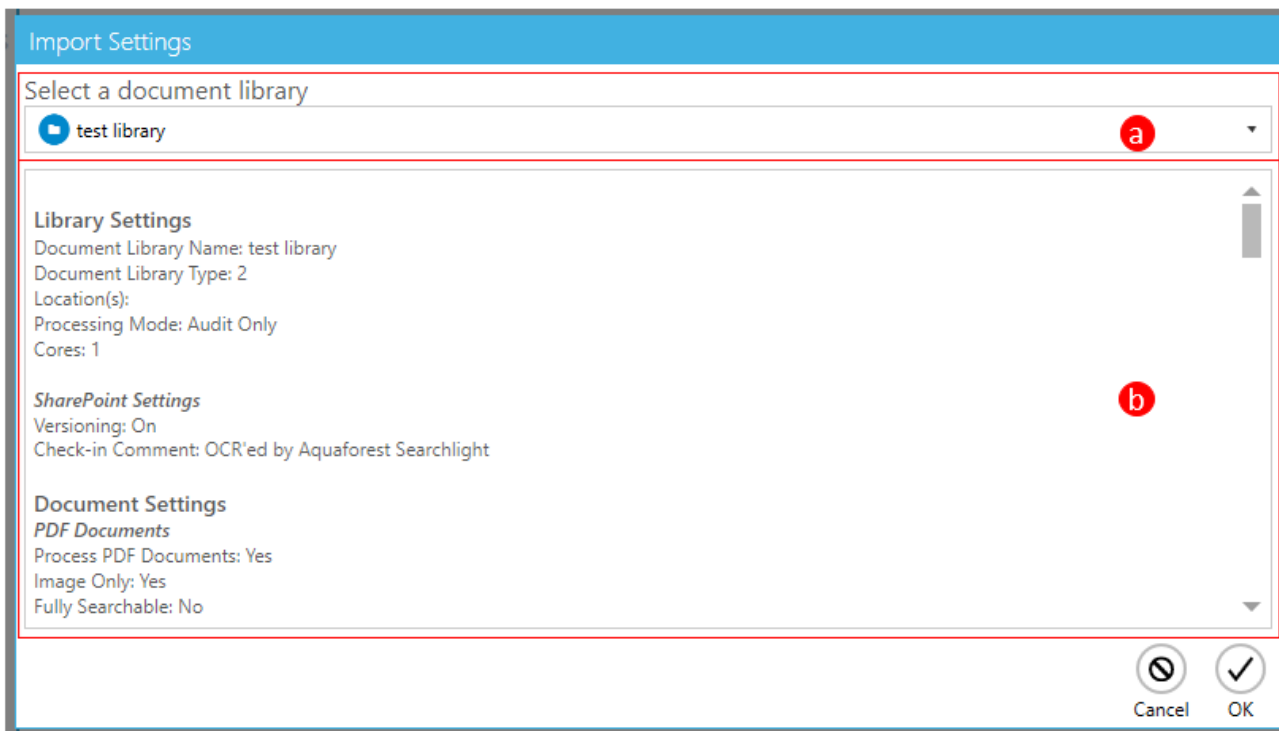
1. Click on the **Add new library** button on the Dashboard to open the wizard.



2. From the wizard select **Import settings from an existing library**



3. From the Import Settings window:
  - a. Select the document library from which the settings are to be imported.
  - b. A summary of the settings of the selected document library will be displayed in the text box underneath.



4. After clicking **OK** from the Import Settings window, go to **Library Settings** and add the location(s) to process. Optionally, add specific locations and documents to exclude.
5. Review all the settings in the other sections and click **Create**.

## 5.4 Audit & Conversion Status

After running a library, its current state will be summarised in the **Statistics** section of the **Status** tab as shown below:

Aquaforest Searchlight

Dashboard Library Settings Help & Support About

Status Library Settings Document Settings Archive Settings OCR Settings Run Details Scheduler Alerts

Big Test - Office365

STATISTICS	LOG OUTPUT
<p><b>PDF Documents</b></p> <ul style="list-style-type: none"> <li>Total PDF Documents: 54</li> <li>Image-only PDFs: 2 (3.8 %)</li> <li>Partially Searchable PDFs: 8 (15.4 %)</li> <li>Fully Searchable PDFs: 42 (80.8 %)</li> <li>Error PDF Documents: 2</li> <li>Total PDF Pages: 984</li> <li>Image-only Pages: 20 (2 %)</li> <li>Fully Searchable Pages: 964 (98 %)</li> </ul> <p><b>Image (TIFF,BMP,JPG,PNG) Documents</b></p> <ul style="list-style-type: none"> <li>Total Image Documents: 9</li> <li>Error Image Documents: 0</li> <li>Total Image Pages: 45</li> </ul> <p><b>MSG Documents</b></p> <ul style="list-style-type: none"> <li>Total MSG Documents: 12</li> <li>Total PDF Attachments: 14</li> <li>Total PDF Pages: 69</li> <li>Image-only Pages: 0 (0 %)</li> <li>Fully Searchable Pages: 69 (100 %)</li> <li>Error MSG Documents: 1</li> </ul> <p><b>Library Totals</b></p>	<p>03-Jan-2020 13:52:55: Finalising conversion</p> <p>Re-calculating statistics after OCR... Document library statistics after OCR:</p> <p>-----</p> <p>PDF Documents</p> <ul style="list-style-type: none"> <li>Total PDF Documents: 54</li> <li>Image-only PDFs: 2 (3.8 %)</li> <li>Partially Searchable PDFs: 8 (15.4 %)</li> <li>Fully Searchable PDFs: 42 (80.8 %)</li> <li>Error PDF Documents: 2</li> <li>Total PDF Pages: 984</li> <li>Image-only Pages: 20 (2 %)</li> <li>Fully Searchable Pages: 964 (98 %)</li> </ul> <p>MSG Documents</p> <ul style="list-style-type: none"> <li>Total MSG Documents: 12</li> <li>Total PDF Attachments: 14</li> <li>Total PDF Pages: 69</li> <li>Image-only Pages: 0 (0 %)</li> <li>Fully Searchable Pages: 69 (100 %)</li> <li>Error MSG Documents: 1</li> </ul> <p>Image (TIFF, BMP, JPEG and/or PNG) Documents</p> <ul style="list-style-type: none"> <li>Total Image Documents: 9</li> <li>Error Image Documents: 0</li> <li>Total Image Pages: 45</li> </ul> <p>Library Totals</p> <ul style="list-style-type: none"> <li>Total Documents: 75</li> <li>Total Error Documents: 3</li> </ul>

Service Status : Running

It provides a breakdown of all the documents processed grouped by the document format. For more detailed analysis of a library, go to the **Run Details** tab.

The screenshot displays the Aquaforest Searchlight interface. At the top, there are navigation tabs: Dashboard, Library, Settings, Help & Support, and About. Below these are sub-tabs: Status, Library Settings, Document Settings, Archive Settings, OCR Settings, Run Details, Scheduler, and Alerts. The main content area is divided into two sections: Run History and Run Details.

**Run History Section:**

- Callout 1: A dropdown menu showing "Showing last 5 runs".
- Callout 2: A table with columns: #, RUN ID, RUN DATE, PROCESSING MODE, STATUS (with a filter icon), AUDIT RESULTS (Successful Documents, Error Documents), STATUS (with a filter icon), and CONVERSION RESULTS (Successful Documents, Error Document).
- Callout 3: Filter icons for the STATUS columns.

**Run Details Section:**

- Callout 2: Radio buttons for "Audit" (selected) and "Conversion".
- Callout 3: A list of document paths.
- Callout 4: A "Limit" dropdown set to "500".
- Callout 5: Navigation arrows for the document list.
- Callout 6: Action buttons: "Export to CSV", "View Full Log", and "Reload".

At the bottom left, the "Service Status" is shown as "Running".

1. Select the number of previous runs to show. You need to click on the **Reload** button after updating this value. Clicking on a run history will display its details in the **Run Details** section below.
2. Select whether you want to display the documents that were audited or OCRed for that specific run.
3. All columns with the ▼ icon next to them can be filtered. You can filter the Searchability status to only display documents that errored during Audit or OCR (Conversion).
4. You can limit the number of documents to display per page. You need to click on the **Reload** button after updating this value.
5. Display the next/previous 500 documents (since **Limit** is set to 500).
6. You can:
  - a. Export the current run details to a CSV file.
  - b. Generate a log file of the current selected run history which will show a file-by-file assessment of all documents processed. The log file can be generated in a PDF, RTF, or HTML format.
  - c. View the log file of the selected run (as displayed in the **Library > Status** tab).



## 6 The Aquaforest Searchlight Tool

### 6.1 Welcome Screen

When Aquaforest Searchlight is launched for the very first time, a Welcome page is displayed to introduce the user to the different features of Aquaforest Searchlight and help in creating the first document library.

**Welcome to Aquaforest Searchlight** Version 2.0 ✕

Aquaforest Searchlight is able to monitor your SharePoint or File System document stores to ensure that all files are fully searchable.

To get started you will need to define a **Searchlight Document Library** that references the document store that you wish to monitor. You can process the Document Library in **Audit Mode** which will scan your documents and provide a report showing the number of files that are not fully searchable.

You can then process the library in **Make Searchable Mode** which will make use of OCR where required to make your documents fully searchable.

There is a sample Searchlight Document Library which you can process to get an understanding of how the product works and the [Reference Guide](#) provides more detailed information.

**Aquaforest**  
© Aquaforest Limited 2001-2020

Show this message on startup  Yes

[Continue](#)





## 6.2 Dashboard

The screenshot shows the Aquaforest Searchlight dashboard. At the top, there are navigation links: Dashboard, Library, Settings, Help & Support, and About. Below the navigation is a '+ Add new library' button. The main content is a table with the following columns: ID, NAME, LIBRARY TYPE, LAST RUN, SCHEDULE, SEARCHABILITY, and RUN STATUS. There are four rows of data, each with a progress bar and control icons (stop, pause, play). The third row is highlighted in blue.

ID	NAME	LIBRARY TYPE	LAST RUN	SCHEDULE	SEARCHABILITY	RUN STATUS
1	Reports (internal)	File System	14-Jan-2020 14:52:00	Manual	100 %	Completed
2	US Patent Front Pages	Office 365	14-Jan-2020 14:37:26	Manual	50 %	Completed
3	Reports (external)	Azure File Share	14-Jan-2020 14:42:28	Manual	94.1 %	Completed
4	Correspondence	Azure Blob Storage	14-Jan-2020 14:56:31	Manual	100 %	Completed

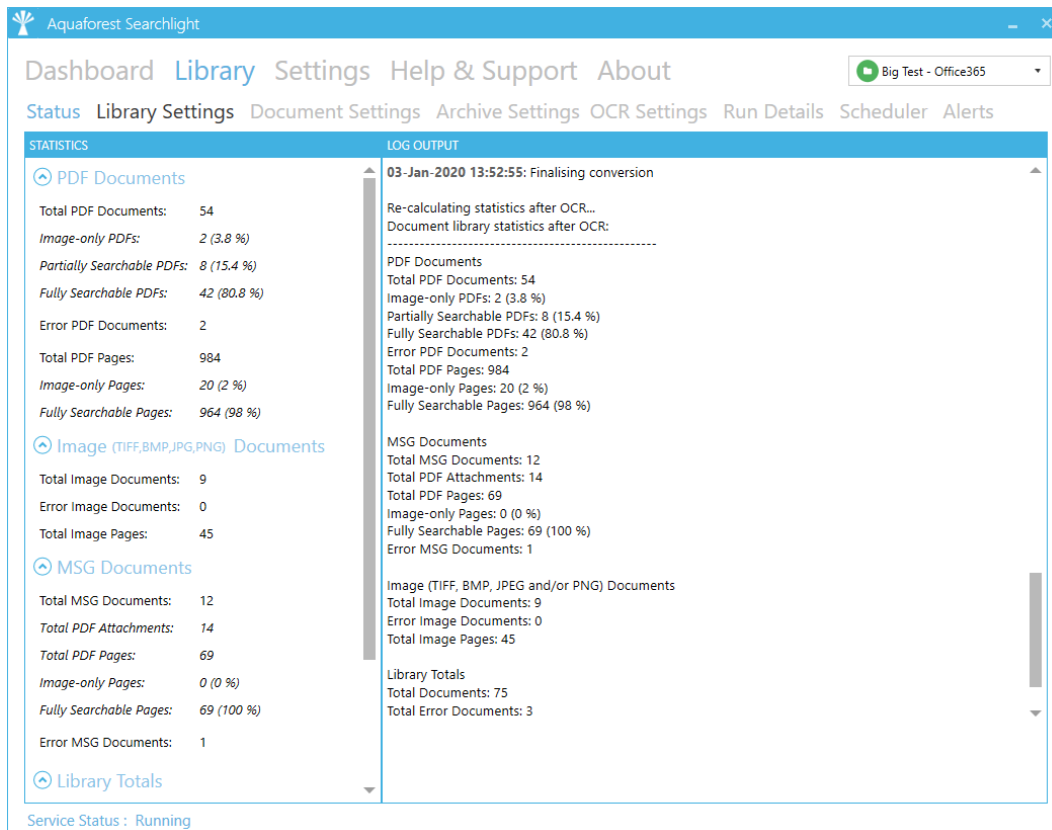
The dashboard gives a summary of the status of all the document libraries that have been created by the user.

Column	Description
Name	Name of the document library
Library Type	The type of the document library: <ul style="list-style-type: none"> <li>• SharePoint On-Premises</li> <li>• SharePoint Online (Office 365)</li> <li>• File System</li> <li>• Azure Blob Storage</li> <li>• Azure File Storage</li> </ul>
Last Run	Time and date of the last run
Schedule	Manual or Automatic
% Searchable	The percentage of pages that is currently searchable in the document library
Run Status	Current status of the document library: <ul style="list-style-type: none"> <li>• Running</li> <li>• Completed</li> <li>• Error</li> <li>• Aborted</li> </ul>
	Abort, Pause, Start

## 6.3 Library

### 6.3.1 Library Status

This screen provides a detailed breakdown of all the document libraries currently configured in Aquaforest Searchlight. Each document library will have detailed information about each of the documents it contains and details about each document.



The screenshot shows the Aquaforest Searchlight interface with the 'Library' tab selected. The 'STATISTICS' section on the left provides a breakdown of document counts and page counts for PDFs, Images, and MSGs. The 'LOG OUTPUT' section on the right shows a log entry for a conversion process on 03-Jan-2020, including re-calculated statistics after OCR.

**STATISTICS**

- PDF Documents**
  - Total PDF Documents: 54
  - Image-only PDFs: 2 (3.8 %)
  - Partially Searchable PDFs: 8 (15.4 %)
  - Fully Searchable PDFs: 42 (80.8 %)
  - Error PDF Documents: 2
  - Total PDF Pages: 984
  - Image-only Pages: 20 (2 %)
  - Fully Searchable Pages: 964 (98 %)
- Image (TIFF,BMP,JPG,PNG) Documents**
  - Total Image Documents: 9
  - Error Image Documents: 0
  - Total Image Pages: 45
- MSG Documents**
  - Total MSG Documents: 12
  - Total PDF Attachments: 14
  - Total PDF Pages: 69
  - Image-only Pages: 0 (0 %)
  - Fully Searchable Pages: 69 (100 %)
  - Error MSG Documents: 1
- Library Totals**
  - Total Documents: 75
  - Total Error Documents: 3

**LOG OUTPUT**

03-Jan-2020 13:52:55: Finalising conversion

Re-calculating statistics after OCR...  
Document library statistics after OCR:  
-----  
PDF Documents  
Total PDF Documents: 54  
Image-only PDFs: 2 (3.8 %)  
Partially Searchable PDFs: 8 (15.4 %)  
Fully Searchable PDFs: 42 (80.8 %)  
Error PDF Documents: 2  
Total PDF Pages: 984  
Image-only Pages: 20 (2 %)  
Fully Searchable Pages: 964 (98 %)

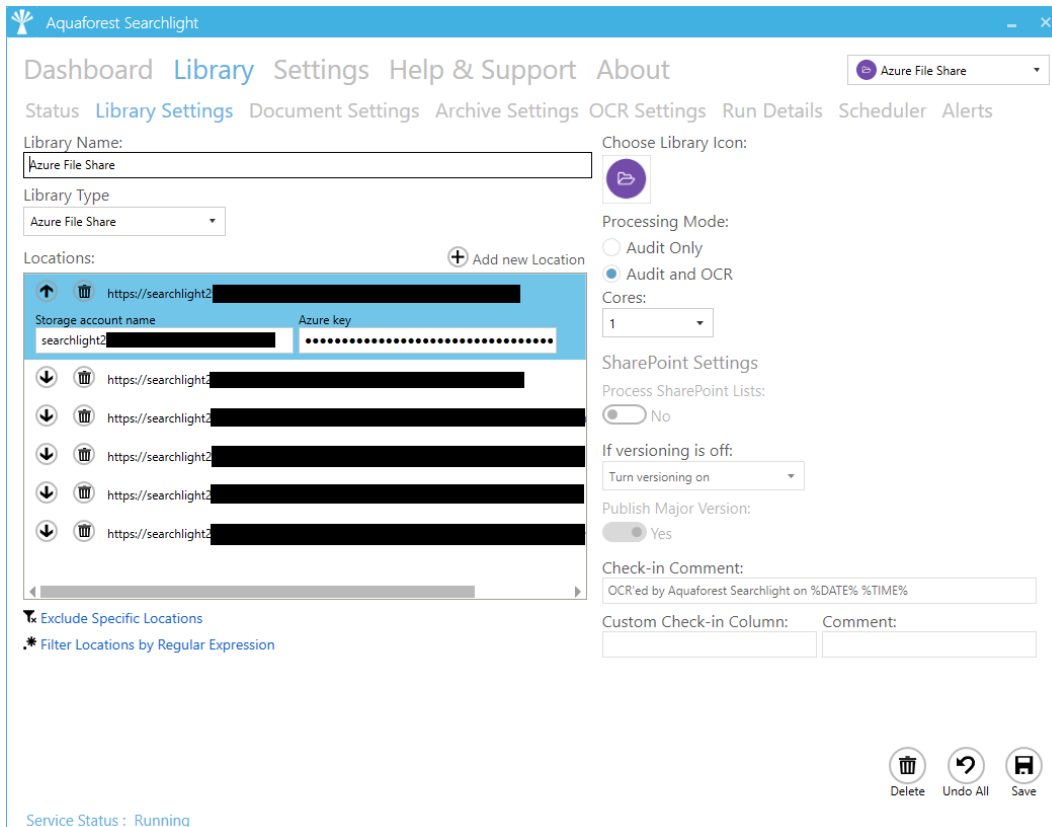
MSG Documents  
Total MSG Documents: 12  
Total PDF Attachments: 14  
Total PDF Pages: 69  
Image-only Pages: 0 (0 %)  
Fully Searchable Pages: 69 (100 %)  
Error MSG Documents: 1

Image (TIFF, BMP, JPEG and/or PNG) Documents  
Total Image Documents: 9  
Error Image Documents: 0  
Total Image Pages: 45

Library Totals  
Total Documents: 75  
Total Error Documents: 3

Service Status : Running

### 6.3.2 Library Settings



The screenshot shows the Aquaforest Searchlight interface with the 'Library Settings' tab selected. The settings are for a library named 'Azure File Share'. The 'Locations' section shows a list of storage locations with their URLs and keys. The 'Processing Mode' is set to 'Audit and OCR'. The 'SharePoint Settings' section includes options for processing SharePoint lists and versioning.

**Library Name:** Azure File Share

**Library Type:** Azure File Share

**Locations:** [+ Add new Location](#)

Storage account name	Azure key
searchlight2	*****
https://searchlight2	
https://searchlight2	
https://searchlight2	
https://searchlight2	
https://searchlight2	

**Choose Library Icon:**

**Processing Mode:**  
 Audit Only  
 Audit and OCR

**Cores:** 1

**SharePoint Settings**  
**Process SharePoint Lists:**  No  
**If versioning is off:** Turn versioning on

**Publish Major Version:**  Yes

**Check-in Comment:** OCR'ed by Aquaforest Searchlight on %DATE% %TIME%

**Custom Check-in Column:**  **Comment:**

Service Status : Running

Delete Undo All Save

Setting	Description
Document Library Name	Name/Title/Description of the document library
Document Library Type	The type of the document library: File System SharePoint Office 365 Azure Blob Storage Azure File Storage
Locations	One or more locations (of the same type) to be processed.
Excluded Specific Locations	Select this if you want to exclude specific locations from being processed. Site collections, sites and libraries that match the specified URLs are excluded.
Filter Locations by Regular Expression	Select this to only include locations whose URLs match specific regular expressions.
Choose Library Icon	Choose an icon to associate to the library.
Processing Mode	<ul style="list-style-type: none"> <li>• Audit Only Analyse the document library to find out the documents that need to be converted without actually converting them.</li> <li>• Audit &amp; OCR Perform audit on the document library and OCR the documents that have been identified as candidates for processing</li> </ul>
Cores	This determines the maximum number of CPU cores that will be used when running the job.
Process SharePoint Lists	Whether or not to process SharePoint lists. NOTE: Process SharePoint lists can be very time consuming if the lists being processed are very large
SharePoint Versioning	This setting can be used to automatically turn versioning on.
Publish Major Version	Publish major version after OCR
Check-in Comment	<p>The check-in comment applied to the updated SharePoint file version.</p> <p>There is also the option of specifying the following templates in the check-in comment:</p> <ul style="list-style-type: none"> <li>• <b>%DATE%</b> : will be replaced by the date the document OCRed</li> <li>• <b>%TIME%</b> : will be replaced by the time the document OCRed</li> </ul>
Custom Check-in Column	Optionally, specify a SharePoint column to add a custom comment to after OCR. NOTE: This is case sensitive.

Setting	Description
Comment	<p>The comment to add to the Custom Check-in Column.</p> <p>There is also the option of specifying the following templates in the comment:</p> <ul style="list-style-type: none"> <li>• <b>%DATE%</b> : will be replaced by the date the document OCR'd</li> <li>• <b>%TIME%</b> : will be replaced by the time the document OCR'd</li> </ul>

### 6.3.3 Document Settings

The screenshot shows the 'Document Settings' window in Aquaforest Searchlight. The settings are organized as follows:

- PDF Selection:** Process PDF Documents (Yes), Image Only PDFs (Yes), Partially Searchable (Yes), Fully Searchable (No), Hidden Text (No).
- BMP Selection:** Process BMP Files (Yes), Delete Original BMP (No).
- JPEG Selection:** Process JPEG Files (Yes), Delete Original JPEG (No).
- PNG Selection:** Process PNG Files (Yes), Delete Original PNG (No).
- MSG Selection:** Process PDF Attachments (Yes).
- Filter Settings:** Temp Folder Location (C:\Aquaforest\Searchlight\temp), Date Filter (No Filter), From/To dates (01/11/2019), Exclude Specific Documents, Filter Documents by Regular Expression.
- Document Error Settings:** Document Error Rule (Copy to Error Folder), Retain Folder Structure (Yes), Error File Template (%FILENAME%%TIMESTAMP%\_%GUID%.%EXT%), Error Location Type (Azure File Share), Document Error Location (https://searchlight2/...).
- Advanced Settings:** Retry On Error (Yes), OCR Document Limit (0), Retain Creation Date (No), Retain Modified Date (No), Retain Created By (No), Retain Modified By (No).

Setting	Description
Process PDF	Whether or not to process PDF documents
Image Only	<p>Whether or not to process Image-only PDFs.</p> <p>An Image-only PDF is a PDF that originated from a scanned document or other digital image. An Image-only PDF does not contain any text, just pictures.</p>

Setting	Description
Partially Searchable	Whether or not to process PDF documents that are partially searchable, i.e., some pages are searchable and some are image-only.
Fully Searchable	Whether or not to process PDF documents that are fully searchable.
Hidden Text	<p>Whether or not process PDF documents with hidden text in them.</p> <p>A Hidden Text PDF has pages that are Image-only with hidden (type 3) text. Such files are typically the output of running an OCR PDF process on an Image Only PDF.</p> <p><b>Note:</b> If you set this setting to true, you might want to consider setting <a href="#">Remove Hidden Text</a> to true in the “OCR Settings &gt; PDF Source Settings”, otherwise you will have multiple OCR text layers per page.</p>
Process TIFF Files	Whether or not to process TIFF files
Delete Original TIFF	Whether or not to delete the original TIFF files after they have been converted to searchable PDFs.
Process BMP Documents	Whether or not to process BMP files.
Delete Original BMP	Whether or not to delete the original BMP files after they have been converted to searchable PDFs.
Process JPEG Files	Whether or not to process JPEG files
Delete Original JPEG	Whether or not to delete the original JPEG files after they have been converted to searchable PDFs.
Process PNG Files	Whether or not to process PNG files.
Delete Original PNG	Whether or not to delete the original PNG files after they have been converted to searchable PDFs.
Process PDF Attachments	Whether or not to process PDF attachments inside MSG files.
Temp Folder Location	The folder used to save documents temporarily for Audit and OCR processing.
Date Filter	Filter out documents by modified or creation date. Documents that fall within the specified “From” and “To” date will be excluded.
Exclude Specific Documents	Select this if you want to exclude specific documents by their paths. Documents that match the specified paths are excluded.
Filter Documents by Regular Expression	Select this to only include documents whose properties match specific regular expressions. E.g., Only include documents whose name matches a specific regular expression.
Document Error Rule	<p>The operation to perform if a document fails to process:</p> <ul style="list-style-type: none"> <li>• Copy to error folder</li> <li>• Move to error folder (for file system library type only)</li> </ul>

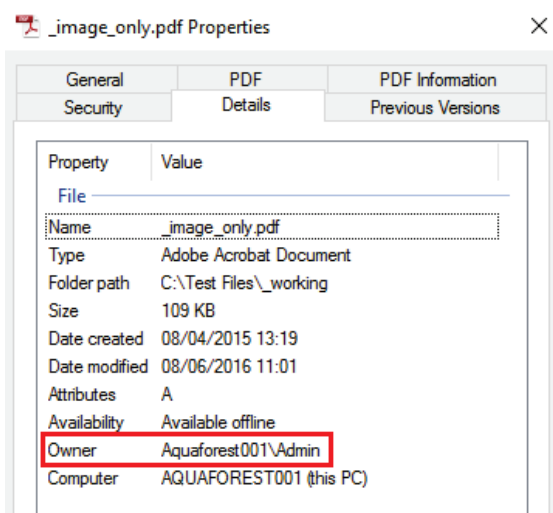
Setting	Description
Retain Folder Structure	Option to retain document's folder structure when copied to error location
Document Error Location	The path of the error location
Document Error Location Type	File System SharePoint Office 365 Azure Blob Storage Azure File Storage
Retry	Whether or not to re-process documents that have previously failed to convert
OCR Document Limit	Limit the number of documents to OCR (not Audit) per run. Set to '0' for no limits.
Retain Creation Date*	Retain the creation date of the source document (SharePoint creation date, FileSystem creation date and created date in PDF properties)
Retain Modified Date*	Retain the modified date of the source document (SharePoint modified date, FileSystem modified date and modified date in PDF properties)
Retain Created By*	Retain the created user of the source document (SharePoint created by FileSystem owner and author in PDF properties)
Retain Modified By*	Retain the created user of the source document (SharePoint modified by)

\* See the sections [6.3.3.1](#), [6.3.3.2](#) and [6.3.3.3](#) for more details about these settings.

### 6.3.3.1 Retain Creation/Modified Date/User

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

- \* "Create User" maps best to "Owner" in Windows File System metadata.



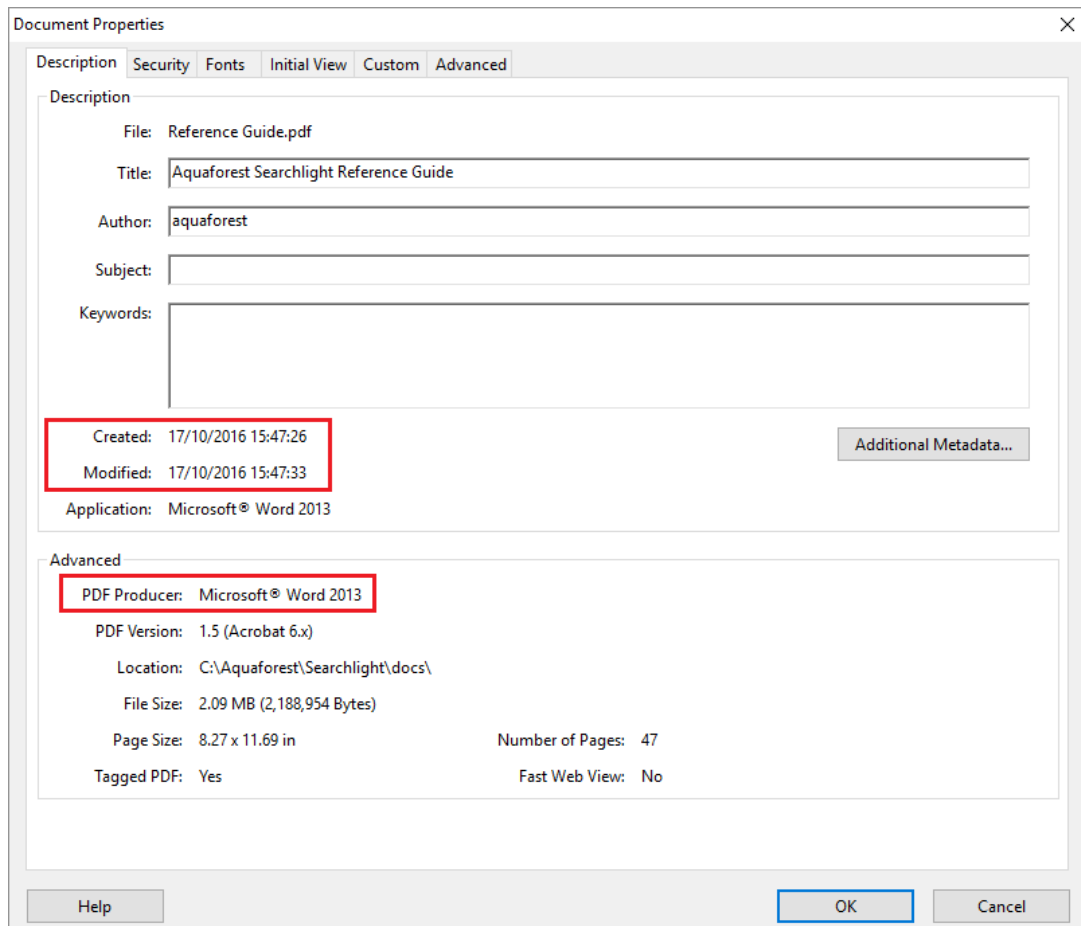
For this to be manipulated, the Searchlight service must be running with sufficient administrative privileges.

- \*\* SharePoint metadata vs. PDF metadata

SharePoint metadata refers to the 'columns' available in SharePoint that stores information about each document.

Column (click to edit)	Type
Title	Single line of text
IM Address	Single line of text
<b>Modified</b>	Date and Time
Created	Date and Time
Created By	Person or Group
<b>Modified By</b>	Person or Group
Checked Out To	Person or Group

PDF metadata refers to the document properties (**File > Properties**) of a PDF document.



### 6.3.3.2 SharePoint Libraries

The behaviour of Retain Creation/Modified Date/User can vary depending on the settings used in SharePoint and Searchlight. The table below summarises when these will and will not be retained in SharePoint.

SharePoint Settings		Searchlight Settings	Created Date retained ?	Created By retained ?	Modified Date retained ?	Modified By retained ?
Create Major Versions	Create Minor Versions	Publish Major Version				
x	x	n/a*	✓	✓	✓	✓
✓	x	n/a*	✓	✓	✓	✓
✓	✓	x	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓	x

n/a\* - To publish major version, *both* major *and* minor versioning must be on in SharePoint.



6.3.3.3 SharePoint Lists

SharePoint Settings	Searchlight Settings	Created Date retained?	Created By retained?	Modified Date retained?	Modified By retained?
Create Versioning	Publish Major Version				
x	n/a	✓	✓	✓	✓
x	n/a	✓	✓	✓	✓
✓	n/a	✓	✓	✓	✓
✓	n/a	✓	✓	✓	✓

## 6.3.4 Document Archive Settings

Aquaforest Searchlight

[Dashboard](#)
[Library](#)
[Settings](#)
[Help & Support](#)

[Status](#)
[Library Settings](#)
[Document Settings](#)
[Archive Settings](#)

### Document Archive Settings

Archive source images to Archive Folder:  
 Yes

Archive source PDF & MSG files to Archive Folder:  
 Yes

Retain folder structure:  
 Yes

Archive Rule:

Archive Template:

Location Type

Archive Location:

[Configure Location](#)

Setting	Description
Archive Template	The template to use to rename the archived file name. The default is: %FILENAME%%TIMESTAMP%.%EXT%
Archive Location	The folder location where original documents will be archived
Archive source Images to Archive folder	If enabled, this will Archive your source Images (TIFF, BMP, JPEG, PNG) to the Archive folder specified above.
Archive source PDF & MSG files to Archive folder	If enabled, this will Archive the source PDFs and MSG files that have PDF attachments to the Archive folder (even when versioning is enabled within SharePoint). A file is only archived before it is OCRed.
Archive Location Type	File System SharePoint Office 365 Azure Blob Storage Azure File Storage
Retain Folder Structure	Option to retain document's folder structure when file is archived

### 6.3.5 OCR Settings

As described in [section 5.1.4](#), Aquaforest Searchlight has 2 OCR engines. When creating a new library, the default OCR settings are loaded from the Properties.xml file for each OCR engine.

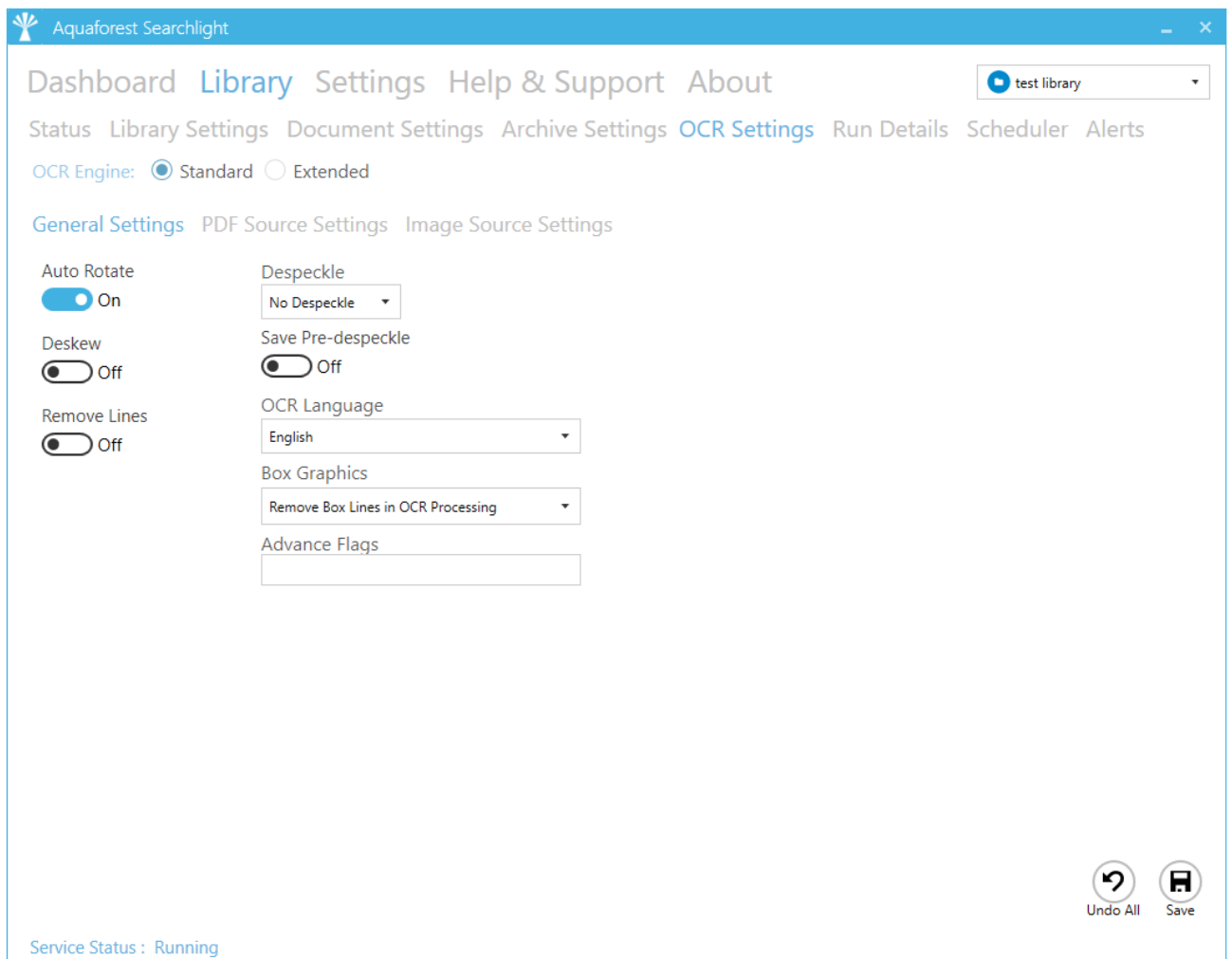
- Aquaforest engine: “[installation path]\ocr\Properties.xml”
- Extended (IRIS) engine: “[installation path]\extendedocr\Properties.xml”

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.

#### 6.3.5.1 Standard OCR Settings

##### 6.3.5.1.1 General Settings



Setting	Description
<b>General Settings</b>	
Auto Rotate	Automatically rotate pages so that text flows left to right
Deskew	Straighten the image

Setting	Description
Remove Lines	Remove lines and boxes during OCR processing to improve recognition – particularly in cases where characters touch lines
Despeckle	Remove specks below the specified pixel size from the image
Box/Graphics Processing	<p>By default, if an area of the document is identified 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.</p> <p>To ensure that the OCR engine can be forced to process such areas there are two options:</p> <p><i>“Treat all Graphics Areas as Text”</i>. This option will ensure the entire document is processed as text.</p> <p><i>“Remove Box Lines in OCR Processing”</i>. 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).</p>
Advanced Flags	Command line flags to be passed through to the underlying executable. Contact <a href="mailto:support@aquaforest.com">support@aquaforest.com</a> for details on using this field.

### 6.3.5.1.2 PDF Source Settings

Aquaforest Searchlight

Dashboard **Library** Settings Help & Support About

Status Library Settings Document Settings Archive Settings **OCR Settings** Run Details Scheduler Alerts

OCR Engine:  Standard  Extended

General Settings **PDF Source Settings** Image Source Settings

Re-Image PDF  No PDF/A  Off

DPI  PDF/A Version

Retain Bookmarks  No Validate PDF/A  Off

Retain Metadata  No

Retain Viewer Prefs  No

Compression  Off

Remove Hidden Text  No

Force Vector Check  No

Undo All Save

Service Status : Running

PDF Source Settings	
Re-Image PDF	Each page of the source PDF is rasterized to an image and appended to a new PDF document.
DPI	Sets the DPI of rasterized images. If 'Re-image PDF' is used, these images will be added to the output file.
Retain Bookmarks	Retains any bookmarks from the source file in the output PDF document when using 'Re-Image PDF'.
Retain Metadata	Retains any metadata from the source file in the output PDF document when using 'Re-Image PDF'.
Retain Viewer Prefs	Retains any PDF Viewer Preferences, Page Mode and Page Layout from source file in the output when using 'Re-Image PDF'
Compression	The image(s) in the output PDF file will be compressed using JBIG2 (for black and white image) or MRC (for color images) which can dramatically reduce the output size of PDFs.
Remove Hidden Text	Remove existing hidden text (text that was added as a result of a previous OCR) from the PDF file so that the resulting searchable PDF file does not have two layers of the same text.

Force Vector Check	This setting is useful when dealing with documents that contains vector objects (e.g., CAD drawings). By default, pages that contain only vector objects are rasterized. Pages that do not have any images but contains vector objects as well as electronic text are skipped from rasterization. However, sometimes there can be a page that contain vector objects (CAD drawings), but its title may be in electronic text. To force rasterizing pages like these, set this property to true.
PDF/A	Switch on to make sure the output PDF conforms to the PDF/A standards.
PDF/A Version	This determines the PDF/A version of the generated PDF.
Validate PDF/A	Validate the PDF as conforming to PDF/A.

### 6.3.5.1.3 Image Source Settings

Aquaforest Searchlight

Dashboard Library Settings Help & Support About

Status Library Settings Document Settings Archive Settings OCR Settings Run Details Scheduler Alerts

OCR Engine:  Standard  Extended

General Settings PDF Source Settings **Image Source Settings**

Compression  Off

PDF/A  Off

PDF/A Version  
A-1b

Validate PDF/A  Off

Undo All Save

Service Status : Running

Image Source Settings	
Compression	The image(s) in the output PDF file will be compressed using JBIG2 (for black and white image) or MRC (for color images) which can dramatically reduce the output size of PDFs.
PDF/A	Switch on to make sure the output PDF conforms to the PDF/A standards.

PDF/A Version

This determines the PDF/A version of the generated PDF.

## 6.3.5.2 Extended OCR Settings

### 6.3.5.2.1 General Settings

Aquaforest Searchlight

Dashboard Library Settings Help & Support About

Status Library Settings Document Settings Archive Settings OCR Settings Run Details Scheduler Alerts

OCR Engine:  Standard  Extended

General Settings PDF Source Settings Image Source Settings Advanced Preprocessing Settings

Auto Rotate  On

Remove Dark Borders  No

Remove Punch Holes  No

Remove Blank Pages  No

Blank Page Sensitivity 1 + -

Work Depth + -

Advanced Flags

Deskew  No

Adjustment Mode

Force Deskew  No

Despeckle  No

Speckle Size No Despeckle

Dilate + -

Remove White Pixels  No

Page Image  Yes

Keep Original Image  Yes

Keep Deskew  No

Keep Despeckle  No

Keep Dark Borders Removal  No

Keep Punch Holes Removal  No

Language(s) Select Language(s)

Bulgarian

Byelorussian

Catalan

Cebuano

Chamorro

Corsican

Croatian

Czech

Danish

Dutch

English

English

Automatic Language Detection  No

Undo All Save

Service Status : Running

Setting	Description
Auto Rotate	Detect page orientation and correct if required
Deskew	Rotates the image to correct its skew angle.
Remove Dark Borders	Removes the dark surrounding from bitonal, grayscale or color images. The dark surrounding of the image is whitened. <b>Note:</b> The dark border should be touching the edge of the image/page for this to work.
Keep Original Image	Yes, to keep the original image as it is. No to output the image generated after selected pre-processing has been applied. <b>Note:</b> This only applies when the source document is an image (TIFF, BMP, JPEG, PNG) or 'Re-Image PDF' is used when the source is a PDF document.

Setting	Description
Despeckle	Removes all the groups of connected pixels with a number of pixels below the parameter.
Advanced Despeckle	The size of the speckles to remove.
Remove White Pixels	By default, despeckle removes black pixels. If set to true, despeckle will remove white pixels rather than black pixels.
Work Depth	This parameter (0 – 255) defines how deeply the OCR engine will analyse a page with 255 being the deepest. For poorer quality documents, higher values can give better recognition results.
Remove Blank Pages	Set this to true to remove blank pages from output PDF documents. A value needs to be set for sensitivity (see below).
Sensitivity	The sensitivity, from 1 to 100. With a high sensitivity, fewer blank pages are detected.
Language	Set the language(s) to use for OCR. <b>Note:</b> <ul style="list-style-type: none"> <li>• Only a maximum of 8 languages can be selected.</li> <li>• Only the English language can be used in conjunction with an Asian language</li> </ul>

### 6.3.5.2.2 PDF Source Settings

The screenshot displays the 'PDF Source Settings' page in the Aquaforest Searchlight application. The page has a blue header with the application name and navigation links for Dashboard, Library, Settings, Help & Support, and About. A dropdown menu for 'SP Online Sites' is visible in the top right. Below the header, there are tabs for Status, Library Settings, Document Settings, Archive Settings, OCR Settings (which is active), Run Details, Scheduler, and Alerts. The 'OCR Engine' is set to 'Extended'. Under 'General Settings', 'PDF Source Settings' is selected, with other options being Image Source Settings and Advanced Preprocessing Settings. The settings are organized into three columns:

- Left Column:**
  - Re-image PDF:  No
  - Output PDF Version: 1.4
  - Validate PDF/A:  No
  - Retain Bookmarks:  No
  - Retain Metadata:  No
  - Retain Viewer Prefs:  No
  - Remove Hidden Text:  No
  - DPI: Auto
  - Force Vector Check:  No
- Middle Column:**
  - Image Compression:  No
  - JPEG Quality: 192
  - JPEG2000 Compression:  No
  - JPEG2000 Compression:  No
  - Compression Mode: Compression Ratio
  - Compression Value: 192
- Right Column:**
  - iHQC Compression:  Off
  - Quality Factor: Medium
  - Compression Level: 1

At the bottom right, there are 'Undo All' and 'Save' buttons. The 'Service Status' at the bottom left is 'Running'.



<b>PDF Source Settings</b>	
Re-Image PDF	Each page of the source PDF is rasterized to an image and appended to a new PDF document.
Output PDF Version	This determines the PDF version of the generated PDF.
Retain Bookmarks	Retains any bookmarks from the source file in the output PDF document when using 'Re-Image PDF'.
Retain Metadata	Retains any metadata from the source file in the output PDF document when using 'Re-Image PDF'.
Remove Hidden Text	Remove existing hidden text (text that was added as a result of a previous OCR) from the PDF file so that the resulting searchable PDF file does not have two layers of the same text.
Remove Visible Text	Whether or not to re-OCR existing visible text.
DPI	Sets the DPI of rasterized images. If 'Re-image PDF' is used, these images will be added to the output file. However, applying 'Image Compression' or 'iHQC Compression' may reduce the DPI in the output PDF.
Force Vector Check	This setting is useful when dealing with documents that contains vector objects (e.g., CAD drawings). By default, pages that contain only vector objects are rasterized. Pages that do not have any images but contains vector objects as well as electronic text are skipped from rasterization. However, sometimes there can be a page that contain vector objects (CAD drawings), but its title may be in electronic text. To force rasterizing pages like these, set this property to true.
Image Compression	Compress color JPEG images in generated PDFs
JPEG Quality	This parameter (0 – 255) determines the compression/quality of color JPEG images in generated PDFs. 0 gives the smallest file size whilst 255 gives the best quality.
JPEG2000 Compression	Use JPEG 2000 compression
Compression Mode	The JPEG 2000 compression mode to use.
Compression Value	The value to use for the selected compression mode.
iHQC Compression	Apply intelligent High-Quality Compression
Quality Factor	The IHQC quality factor.
Compression Level	The iHQC compression level to be used. Level 1 is the basic compression level. Level 3 is the most advanced intelligent High-Quality Compression mode.

### 6.3.5.2.3 Image Source Settings

Aquaforest Searchlight

Dashboard **Library** Settings Help & Support About

Status Library Settings Document Settings Archive Settings **OCR Settings** Run Details Scheduler Alerts

OCR Engine:  Standard  Extended

General Settings PDF Source Settings **Image Source Settings** Advanced Preprocessing Settings

Output PDF Version: 1.4

Validate PDF/A:  No

**Image Compression**

Image Compression:  No

JPEG Quality: 192

**JPEG2000 Compression**

JPEG2000 Compression:  No

Compression Mode: Quality Factor

Compression Value: 192

**iHQC Compression**

iHQC Compression:  Off

Quality Factor: Medium

Compression Level: 1

Undo All Save

Service Status: Running

Image Source Settings	
Output PDF Version	This determines the PDF version of the generated PDF.
Image Compression	Compress color JPEG images in generated PDFs
JPEG Quality	This parameter (0 – 255) determines the compression/quality of color JPEG images in generated PDFs. 0 gives the smallest file size whilst 255 gives the best quality.
JPEG2000 Compression	Use JPEG 2000 compression
Compression Mode	The JPEG 2000 compression mode to use.
Compression Value	The value to use for the selected compression mode.
iHQC Compression	Apply intelligent High-Quality Compression
Quality Factor	The IHQC quality factor.
Compression Level	The iHQC compression level to be used. Level 1 is the basic compression level. Level 3 is the most advanced intelligent High-Quality Compression mode.

### 6.3.5.2.4 Advanced Pre-processing Settings

**Line Removal**

Remove Lines  No

Clean X	Horizontal 1 + -	Vertical 1 + -
Clean Y	Horizontal 1 + -	Vertical 1 + -
Dilate	Horizontal 1 + -	Vertical 1 + -
Max Gap	Horizontal 2 + -	Vertical 2 + -
Max Thickness	Horizontal 16 + -	Vertical 16 + -
Min Length	Horizontal 128 + -	Vertical 128 + -

**Binarization**

Binarize  No

Binarization Mode [Dropdown]

Threshold 128 + -

Contrast 40 + -

Brightness 0 + -

Smoothing Level 0 + -

Undithering  No

**Interpolation**

Interpolate  No

Interpolation Mode [Normal] [Dropdown]

Interpolation Value + -

Undo All Save

Service Status : Running

Advanced Pre-processing Settings	
Remove Lines	Whether or not to remove lines from an image (The image must be black and white).
Horizontal Clean X	The parameter for cleaning noisy pixels attached to the horizontal lines.
Horizontal Clean Y	The parameter for cleaning noisy pixels attached to the horizontal lines.
Vertical Clean X	The parameter for cleaning noisy pixels attached to the vertical lines.
Vertical Clean Y	The parameter for cleaning noisy pixels attached to the vertical lines.
Horizontal Dilate	The dilate parameter that helps the detection of horizontal lines.
Vertical Dilate	The dilate parameter that helps the detection of vertical lines.
Horizontal Max Gap	The maximum horizontal line gap to close. It is useful to remove broken lines.
Vertical Max Gap	The maximum vertical line gap to close. It is useful to remove broken lines.

Horizontal Max Thickness	The maximum thickness of the horizontal lines to remove. It is useful to keep vertical lines larger than this parameter. Can be also useful to keep vertical letter strokes.
Vertical Max Thickness	The maximum thickness of the vertical lines to remove. It is useful to keep horizontal lines larger than this parameter. Can be also useful to keep horizontal letter strokes.
Horizontal Min Length	The minimum length of the horizontal lines to remove.
Vertical Min Length	The minimum length of the vertical lines to remove.
Binarize	Whether or not to perform binarization on the document.
Brightness	The brightness (higher values will darker the result).
Contrast	The contrast (lower values will darker the result).
Smoothing Level	Smoothing may be useful to binarize text with a colored background in order to avoid noisy pixels (0 disables smoothing, higher values smooth more).
Threshold	Sets the threshold for fixed threshold binarization (0 for automatic threshold computation).
Interpolate	Whether or not to interpolate.
Interpolation Mode	Sets the interpolation mode.
Interpolation Value	Interpolates the source image to the given resolution. This value (the target resolution) must be greater than the source image's resolution.

## 6.3.6 Run Details

Aquaforest Searchlight

Dashboard Library Settings Help & Support About

Reports (external)

Status Library Settings Document Settings Archive Settings OCR Settings Run Details Scheduler Alerts

Run History Showing last 5 runs

#	RUN ID	RUN DATE	PROCESSING MODE	Status	AUDIT RESULTS		Status	CONVERSION RESULTS	
					Successful Documents	Error Documents		Successful Documents	Error Documents
1	2	14-Jan-2020 14:42:28	Audit and OCR	Completed	61	3	Completed	1	1

Run Details

Audit Conversion

#	DOCUMENT PATH	SEARCHABILITY	FILE TYPE
1	https://[REDACTED].PDFs/[REDACTED].pdf	searchable	.PDF
2	https://[REDACTED].PDFs/[REDACTED].pdf	searchable	.PDF
3	https://[REDACTED].PDFs/[REDACTED].pdf	searchable	.PDF
4	https://[REDACTED].PDFs/[REDACTED].pdf	searchable	.PDF
5	https://[REDACTED].PDFs/[REDACTED].pdf	searchable	.PDF
6	https://[REDACTED].PDFs/[REDACTED].pdf	searchable	.PDF
7	https://[REDACTED].PDFs/[REDACTED].pdf	searchable	.PDF
8	https://[REDACTED].PDFs/[REDACTED].pdf	searchable	.PDF
9	https://[REDACTED].PDFs/[REDACTED].pdf	searchable	.PDF

Limit 500

Export to CSV View Full Log Reload

Service Status : Stopped

Previous runs carried out on a particular document library are listed under the **Run History** section. The **Run Details** list provide detailed information about each run. Both the Run History and Run Details have columns where filters can be applied to limit what is displayed.

Use **Export to CSV** to export the run details to CSV file.

The **View Full Log** button can be used to display the full log file of a specific run.

### 6.3.6.1 Run Details Context Menu

Use the right-click context menu to:

- Copy the file path of the selected document.
- Open the file (File System and SharePoint only)
- Open the location of the file (File System and SharePoint only)

Aquaforest Searchlight

Dashboard [Library](#) [Settings](#) [Help & Support](#) [About](#) Azure File Share

Status [Library Settings](#) [Document Settings](#) [Archive Settings](#) [OCR Settings](#) [Run Details](#) [Scheduler](#) [Alerts](#)

Run History Showing last 5 runs

#	RUN ID	RUN DATE	PROCESSING MODE	Status	AUDIT RESULTS		Status	CONVERSION RESULTS	
					Successful Documents	Error Documents		Successful Documents	Error Documents
1	16505	14-Jan-2020 13:55:01	Audit Only	Completed	1	0	No Conversions	0	0
2	16504	14-Jan-2020 13:53:23	Audit Only	Error	0	0	No Conversions	0	0
3	16504	14-Jan-2020 13:37:29	Audit Only	Completed	0	0	No Conversions	0	0

Run Details  Audit  Conversion

#	DOCUMENT PATH	SEARCHABILITY	FILE TYPE	LAST MODIFIED	PAGES
1	https://.../source-documents/Test Files/6/_image_only.pdf	imageonly	.PDF	14-Jan-2020 13:54:54	1

Context menu for document path:

- Copy File Path
- Open File
- Open File Location

### 6.3.7 Scheduler Settings

Aquaforest Searchlight

Dashboard [Library](#) [Settings](#) [Help & Support](#) [About](#) Office365

Status [Library Settings](#) [Document Settings](#) [Archive Settings](#) [OCR Settings](#) [Run Details](#) [Scheduler](#) [Alerts](#)

Manual

Once per day

At:

Continuous

Every:  Hour(s)

Between  And

Run once

On:

At:

Undo All Save

Service Status : Running

Setting	Description
Manual	This means that the document library must be run manually by clicking on the "Run" button on the dashboard.

Setting	Description
Once per day	This allows the document library to be scheduled to run at a specified time each day.
Continuous	This allows the document library to be scheduled to run periodically between a start time and end time each day. The periods may be minutes, hours, days, or months. For example, a document library may be specified to run every 1 hour between 9:00 and 17:00.
Run Once	This allows the document library to be scheduled to run only once at a specified time.

## 6.3.8 Alert Settings

### 6.3.8.1 Action

The screenshot shows the 'Alerts' configuration page in Aquaforest Searchlight. The page has a blue header with the application name and navigation links: Dashboard, Library, Settings, Help & Support, and About. A dropdown menu shows 'test library'. Below the header are tabs for Status, Library Settings, Document Settings, Archive Settings, OCR Settings, Run Details, Scheduler, and Alerts. The 'Alerts' tab is active, and the 'Action' configuration section is expanded. The 'Action' section asks 'What action(s) do you want the alert task to perform?' and includes the following settings:

- Action:** What action(s) do you want the alert task to perform?
- Email:** Send an email (toggle: Yes)
- Report:** Generate a CSV report (toggle: Yes)
- Trigger:** Attach the CSV report to the email (toggle: No)
- Finish:** Save Report (toggle: Yes)
- Location Type:** File System (dropdown menu)
- Location:** C:\sl2test\Reports (text input field)

At the bottom right, there are navigation buttons: Previous, Next, View Alert Log, Undo All, and Save. The service status is shown as 'Running' at the bottom left.

Setting	Description
<b>Action</b>	
Send an email	Select this if you want to send an email
Generate a CSV report	Select this if you want to generate a report
Attach the CSV report to the email	Whether or not to attach the CSV report to the email

Setting	Description
Save Report	Save the report locally
Location Type	The type of storage used to save the report: File System SharePoint Office 365 Azure Blob Storage Azure File Storage
Location	The location to save the report

### 6.3.8.2 Email

The screenshot shows the Aquaforest Searchlight web interface. The top navigation bar includes 'Dashboard', 'Library', 'Settings', 'Help & Support', and 'About'. A dropdown menu is set to 'File System'. Below the navigation, there are tabs for 'Status', 'Library Settings', 'Document Settings', 'Archive Settings', 'OCR Settings', 'Run Details', 'Scheduler', and 'Alerts'. The main content area is titled 'Configuration Action > Email'. On the left, there is a sidebar with 'Action', 'Email', 'Report', 'Trigger', and 'Finish'. The main area contains the following fields:

- From:** name.surname@companyxyz.com
- To:** name.surname@companyxyz.com
- Cc:** (empty)
- Bcc:** admin@companyxyz.com
- Subject:** %LIBRARYNAME% - %STATUS%
- Message:** Library Name: %LIBRARYNAME%  
Job Status: %STATUS%  
Log File Path: %LOGFILEPATH%

At the bottom right, there are buttons for 'Test Email', 'Previous', 'Next', 'View Alert Log', 'Undo All', and 'Save'. The service status at the bottom left is 'Stopped'.

Email	
From	The email address to send the email from.
To Cc Bcc	The email address(es) to send the email to. Multiple email addresses can be specified by separating each one with a semicolon in the "To", "Cc" and "Bcc" fields.



Subject	<p>The email subject. You can use the following templates:</p> <ul style="list-style-type: none"> <li>• <b>%LIBRARYNAME%</b> - will be replaced by the name of the library</li> <li>• <b>%STATUS%</b> - will be replaced by "success" or "error" depending on whether the job ran successfully or not</li> </ul>
Message	<p>The email message to send. You can use the following templates within the email message:</p> <ul style="list-style-type: none"> <li>• <b>%LIBRARYNAME%</b> - will be replaced by the name of the library</li> <li>• <b>%STATUS%</b> - will be replaced by "success" or "error" depending on whether the job ran successfully or not</li> <li>• <b>%LOGFILEPATH%</b> - will be replaced by the path of the log file for the library</li> <li>• <b>%ERRORMESSAGE%</b> - will be replaced by any error messages that occurred during the library run</li> </ul>

### 6.3.8.3 Report

The screenshot shows the Aquaforest Searchlight web interface. The top navigation bar includes 'Dashboard', 'Library', 'Settings', 'Help & Support', and 'About'. A dropdown menu is set to 'test library'. Below the navigation, there are tabs for 'Configuration' and 'Action > Report'. The 'Report' configuration page is active, showing options for 'Library Audit Summary' and 'Run Details Summary (OCR only)'. The 'Library Audit Summary' section has a 'Show library audit summary in report' toggle set to 'Yes'. The 'Run Details Summary (OCR only)' section has three toggles: 'Show run details summary in report' (set to 'No'), 'Show details of individual documents that were processed' (set to 'No'), and 'Choose the columns that will appear in the report' (with 'Document Path' and 'Searchability' toggles set to 'No'). At the bottom right, there are navigation buttons for 'Previous' and 'Next', and action buttons for 'View Alert Log', 'Undo All', and 'Save'. The service status at the bottom left is 'Running'.

## Report

Show library audit summary in report	The library audit summary contains statistics about the current searchability status of the library as a whole, as well as individual statistics about each document type in the library.
Run Details Summary (OCR only)	
Show run details summary in report.	The run details summary lists all the documents that were processed in a particular run including: <ul style="list-style-type: none"> <li>• Number of documents OCR'd</li> <li>• Number of documents that failed to OCR</li> </ul>
Show details of individual documents that were processed	Include in the report individual document details (for the columns to be included see below)
Limit	Set the maximum number of documents reported. This value needs to be set by the user.
Choose columns that will appear in the report:	The columns include: Document Path Searchability Document Type Number of pages Number of searchable pages Number of image pages Conversion status

#### 6.3.8.4 Trigger

Aquaforest Searchlight

Dashboard Library Settings Help & Support About

Status Library Settings Document Settings Archive Settings OCR Settings Run Details Scheduler Alerts

test library

Configuration Trigger

Action

Email

Report

Trigger

Finish

When do you want the alert task to run?

Every time the library runs successfully

No

Every time the library fails to run

No

Every time there is a SharePoint or Azure connection error

No

Advanced Settings

Independent of the above trigger settings, also run the alert task on a schedule

Yes

At 11:15.

Start: 27/08/2019 11:15

Every: 1 day(s)

Daily

Weekly

Monthly

One time

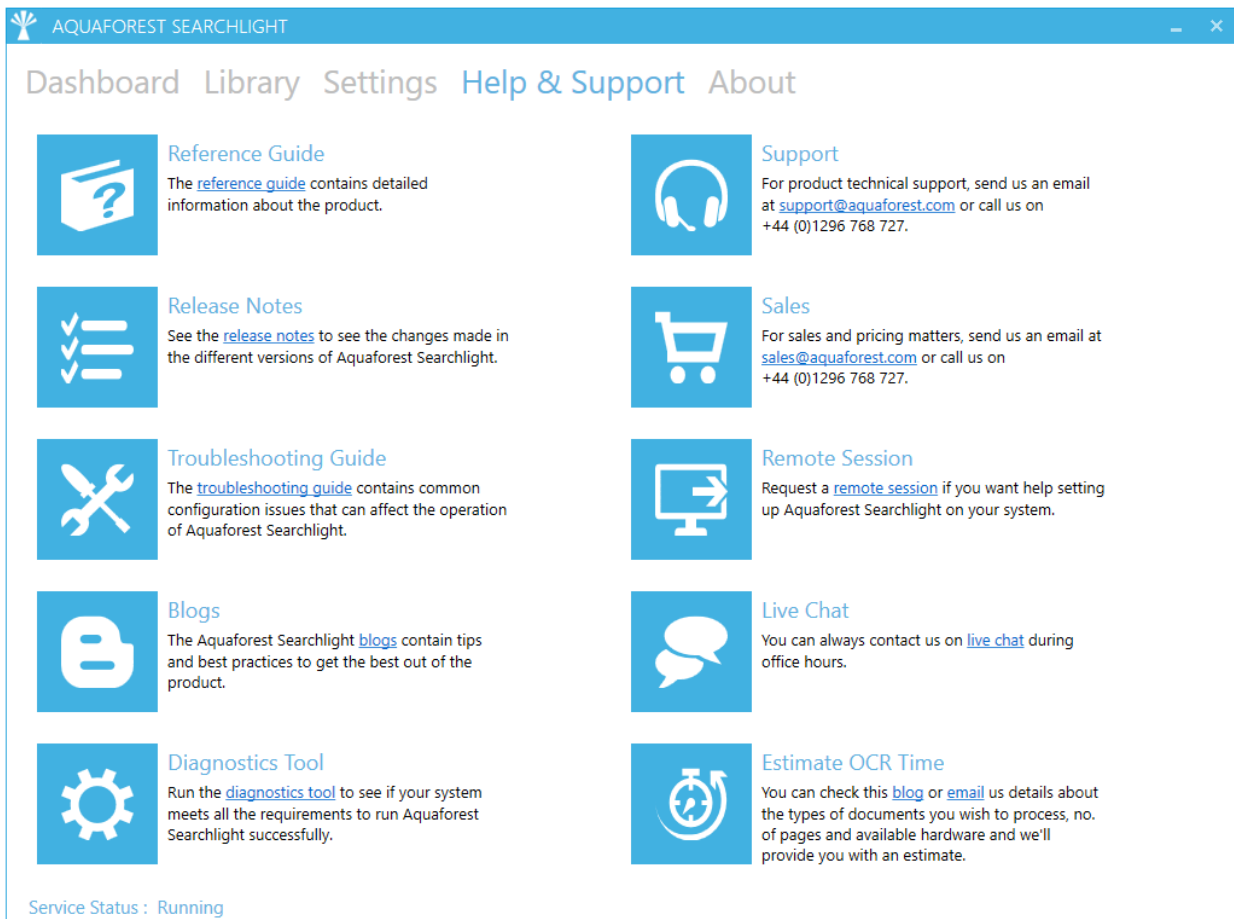
Previous Next

View Alert Log Undo All Save

Service Status : Running

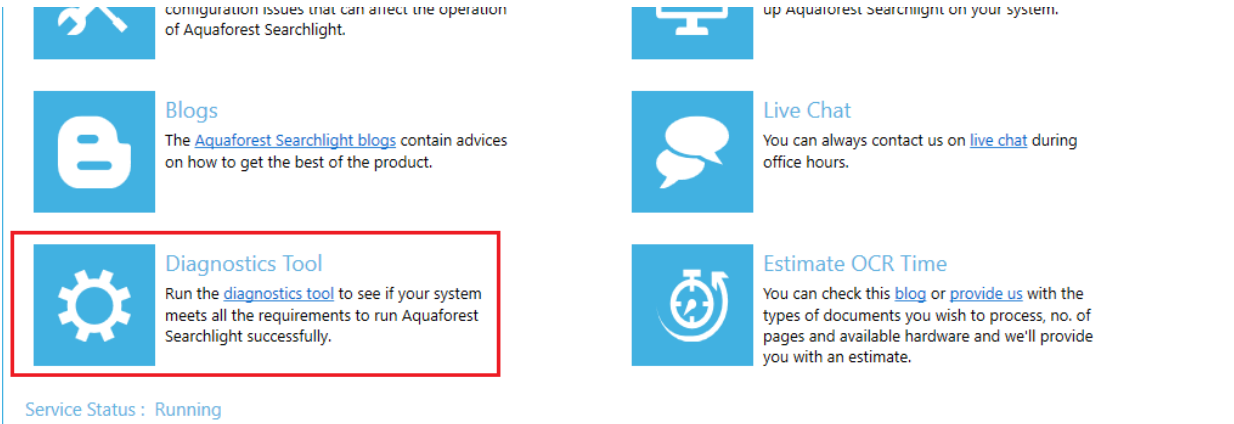
Trigger	
Alert is triggered	Every time the library runs successfully. Every time the library fails to run. Every time there is a SharePoint or Azure connection error
Advanced Settings	Independent of the above trigger settings, the alert can be scheduled to run daily, weekly (on selected days), monthly or once.
Expires	Whether or not the trigger expires
Expiry	The expiry date of the trigger. The alert task will not run after this date.

## 6.4 Help & Support



The Help & Support page is the starting point for help with Aquaforest Searchlight. It provides resources such as the reference guide, release notes and online blogs. It also provides the generic support email address which should be used in the first instance when reporting an issue or any queries.

### 6.4.1 Diagnostic Tool



To run the diagnostic tool, click on the “Diagnostics Tool” icon in the “Help & Support” tab as highlighted in the image above. This will initiate the diagnostic wizard which will run various checks to determine if your system meets all the requirements needed to run Aquaforest Searchlight as well as collect information related to a specific document library. All the gathered information will be made available in a zip file which can be sent to [support@aquaforest.com](mailto:support@aquaforest.com) for further investigation.

### 6.4.2 Database Clean-up Tool

Running Searchlight over a long period of time can dramatically increase the database size. This can be an issue if space is limited in the server running Searchlight.

Searchlight comes with a command line tool that will try to compact the database by deleting logs from previous runs.

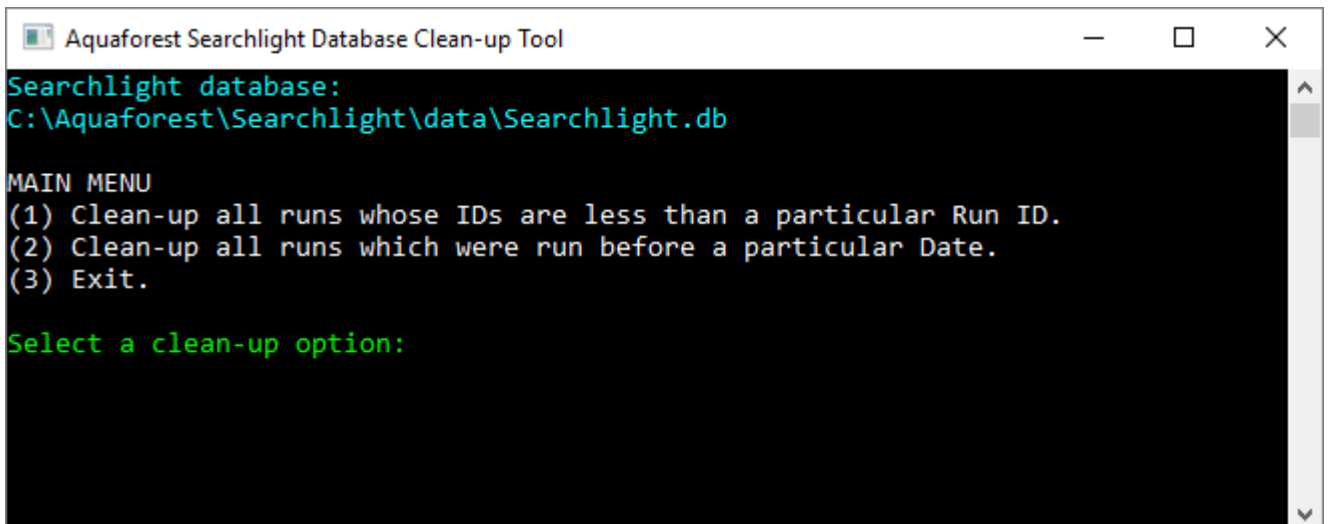
The clean-up tool is located at “[Install location]/bin/ Aquaforest.Searchlight.DatabaseCleanup.exe”.

The runs from which the logs are to be deleted can be selected either by date last run or by the Run ID. This information can be obtained from the Dashboard by selecting a document library that had been run recently and going to the status tab.

[Status](#) [Library Settings](#) [Document Settings](#) [Archive Settings](#) [OCR Se](#)

STATISTICS	LOG OUTPUT
<p>PDF Documents</p> <p>Total PDF Documents: 22</p> <p>Image-only PDFs: 1 (5 %)</p> <p>Partially Searchable PDFs: 0 (0 %)</p> <p>Fully Searchable PDFs: 19 (95 %)</p> <p>Error PDF Documents: 2</p> <p>Total PDF Pages: 167</p> <p>Image-only Pages: 1 (0.6 %)</p> <p>Fullv Searchable Paaes: 166 (99.4 %)</p>	<p>Aquaforest Searchlight 2.0.191206.0</p> <p>Document Library ID: 2   Run ID: 4</p> <p>31-Dec-2019 10:16:29: Job Start</p> <p>31-Dec-2019 10:16:29: Starting Audit...</p> <p>Enumerating documents...</p> <p>Checking library <a href="https://aquaforest.sharepoint">https://aquaforest.sharepoint</a></p> <p>Checking library <a href="https://aquaforest.sharepoint">https://aquaforest.sharepoint</a></p> <p>Checking library <a href="https://aquaforest.sharepoint">https://aquaforest.sharepoint</a></p> <p>Checking library <a href="https://aquaforest.sharepoint">https://aquaforest.sharepoint</a></p> <p>Checking library <a href="https://aquaforest.sharepoint">https://aquaforest.sharepoint</a></p> <p>Checking library <a href="https://aquaforest.sharepoint">https://aquaforest.sharepoint</a></p> <p>Checking library <a href="https://aquaforest.sharepoint">https://aquaforest.sharepoint</a></p> <p>Documents enumerated (matching selection  </p>

With this information, log in as an Administrator and start the command line tool.



```
Aquaforest Searchlight Database Clean-up Tool
Searchlight database:
C:\Aquaforest\Searchlight\data\Searchlight.db

MAIN MENU
(1) Clean-up all runs whose IDs are less than a particular Run ID.
(2) Clean-up all runs which were run before a particular Date.
(3) Exit.

Select a clean-up option:
```

## 6.5 Settings

### 6.5.1 License Settings

Aquaforest Searchlight

Dashboard Library **Settings** Help & Support About

License Email Theme Date & Time Advanced

License Type: Permanent

Computer Bound: No

Multi-core: No

Max Cores: 1

Document Limit: Unlimited

Trial Stamp: No

Expires: No

Library Type(s): File System: ✓ | SharePoint: ✓ | Office 365: ✓ | Azure: ✓

OCR Features: Standard OCR: ✓ | Extended OCR: ✓ (Asian Languages: ✓; Arabic & Farsi Languages: ✓; Hebrew Language: ✓; IHQC: ✓)

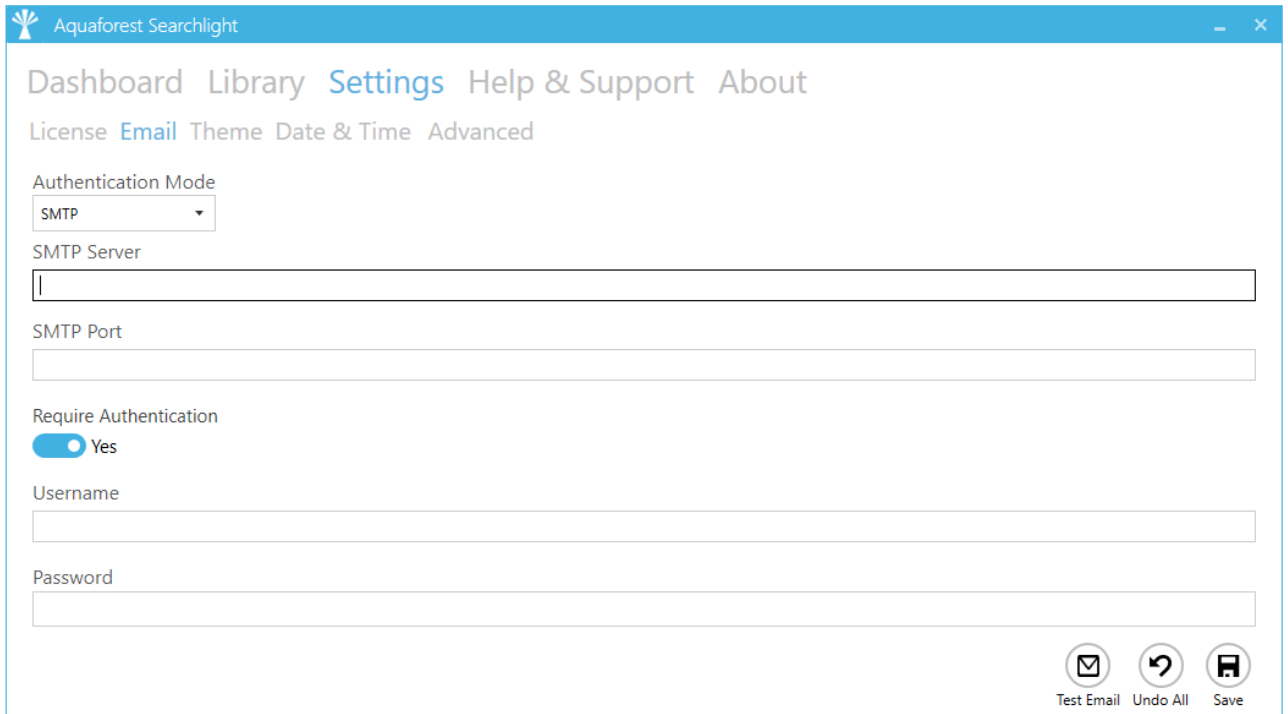
License Key:

Setting	Description
License Type	Trial or Permanent
Computer Bound	Whether the license is computer bound or not computer bound
Computer Identifier	The name of the computer if the license is computer bound
Multi-core	Whether or not the license allows the use of multiple cores for processing
Max Cores	The maximum number of cores that can be used for processing
Document Limit	The maximum number of documents that can be OCR'd. If this limit is reached, OCR will be disabled.
Trial Stamp	Whether or not the OCR'd documents will have a trial stamp
Expires	Whether the license has an expiry date
Features	Modules enabled by the current license
License Key	The license key currently being used

## 6.5.2 Email Settings

The Email tab allows email server information to be defined. This is used to support the “[Email Alerts](#)” functionality. Aquaforest Searchlight supports two authentication modes: SMTP and Azure OAuth2.

### 6.5.2.1 SMTP



The screenshot shows the Aquaforest Searchlight interface with the 'Settings' tab selected. Under the 'Email' sub-tab, the 'Authentication Mode' is set to 'SMTP'. The 'SMTP Server' field is empty. The 'SMTP Port' field is also empty. The 'Require Authentication' toggle is turned on to 'Yes'. The 'Username' and 'Password' fields are empty. At the bottom right, there are three icons: 'Test Email', 'Undo All', and 'Save'.

Setting	Description
SMTP Server	Address of the server hosting the SMTP server.
SMTP Port	SMTP Server port. Standard SMTP ports: 25, 587 or 465
Require Authentication	The email address used for the sender must be authenticated using the username and password below.
Username	Username for authentication by the server.
Password	Password for the username.

## 6.5.2.2 Azure OAuth2

For additional details on OAuth2 authentication, refer to “**Exchange Online OAuth2 Configuration.pdf**” document in the **docs** folder where Searchlight is installed.

The screenshot shows the Aquaforest Searchlight settings interface. At the top, there are navigation links: Dashboard, Library, Settings (highlighted), Help & Support, and About. Below these are sub-links: License, Email, Theme, Date & Time, and Advanced. The main settings area is titled 'Authentication Mode' and has a dropdown menu set to 'OAuth2 (Azure)'. Below this are four text input fields: 'Azure Client ID', 'Azure Tenant', 'Azure AD Instance' (with a hint 'E.g. https://login.microsoftonline.com'), and 'Client Secret'. A 'Credential Type' dropdown is set to 'Client secret'. At the bottom right of this section are three icons: 'Test Email', 'Undo All', and 'Save'.

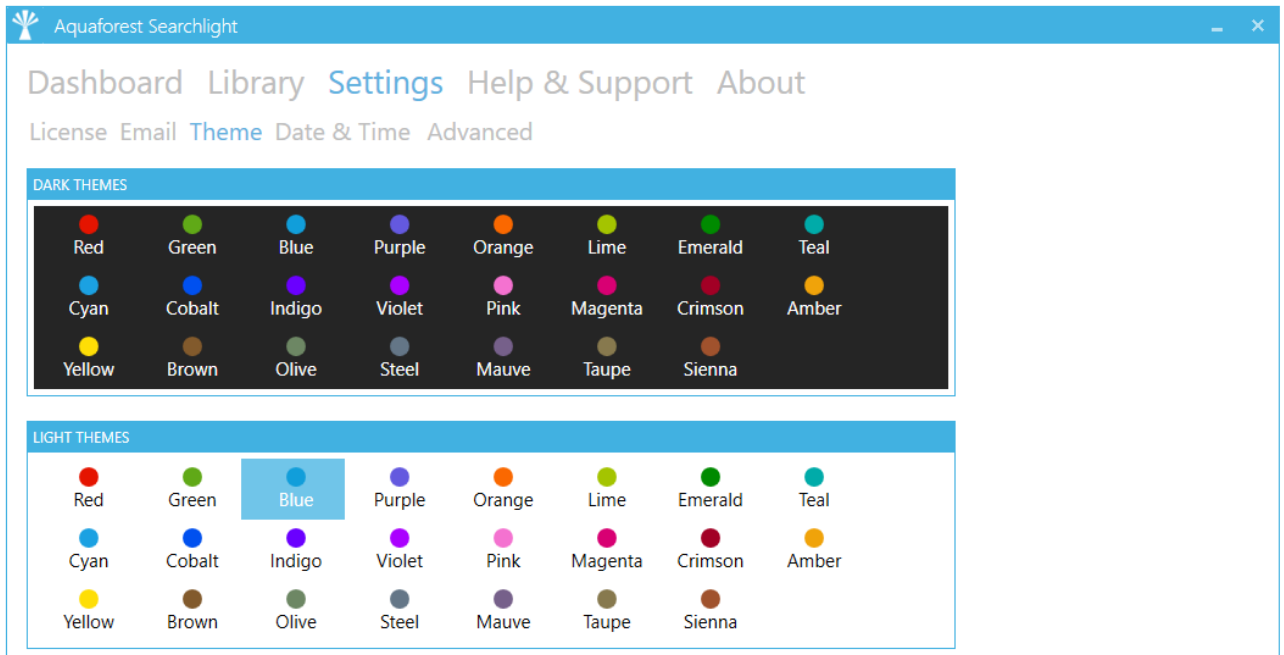
This screenshot shows the 'Certificate' authentication settings section. It features a 'Credential Type' dropdown set to 'Certificate'. Below it are two text input fields: 'Certificate Path' and 'Certificate Password'. At the bottom right, there are three icons: 'Test Email', 'Undo All', and 'Save'.

Setting	Description
Azure Client ID	The Application GUID used by the application to uniquely identify itself to Azure AD
Azure Tenant	The tenant ID of the Azure AD tenant in which this application is registered (a GUID)
Azure AD Instance	Instance of Azure AD, for example public Azure or a Sovereign cloud (Azure China, Germany, US government, etc...)  The default value is: <a href="https://login.microsoftonline.com">https://login.microsoftonline.com</a>
Client Secret	The client secret to use to access the Azure application
Certificate Path	The local path of the certificate previously shared with Azure AD during the application registration
Certificate Password	The password for the certificate



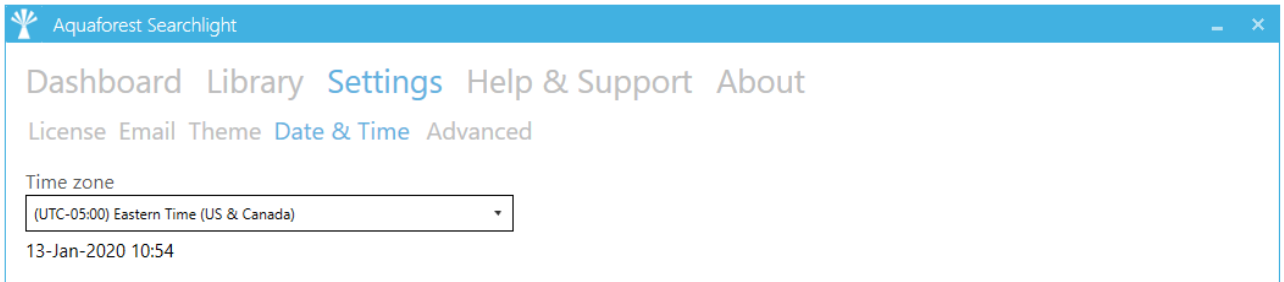
### 6.5.3 Themes

There is a selection of 23 accent colors available split between dark and light themes. The Light Blue is the default theme.



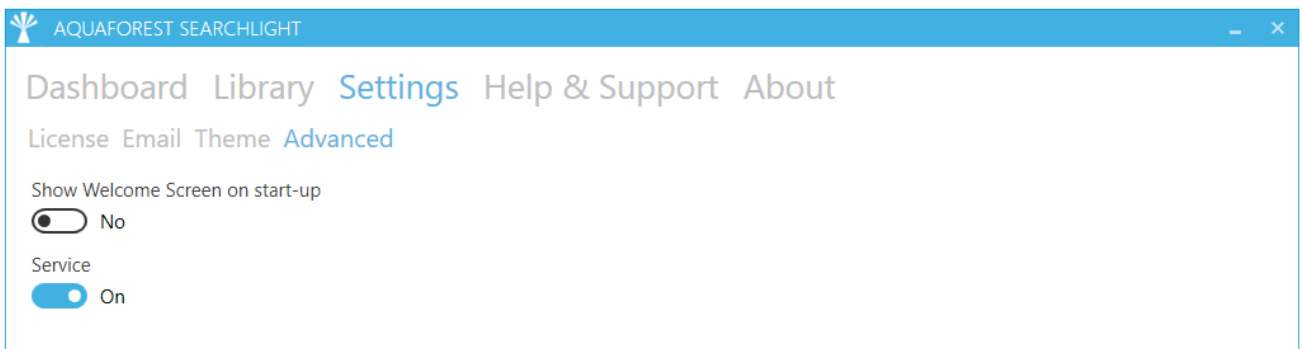
### 6.5.4 Date & Time

Set time zone (relative to UTC).



Internally Searchlight 2.0 uses UTC dates and times, any local file date and times are converted using the selected time zone settings to UTC.

### 6.5.5 Advanced Settings



Setting	Description
Show Welcome Screen on start-up	Whether or not to show the <a href="#">Welcome Screen</a> when launching the Aquaforest Searchlight UI.
Service	Switch to turn the Aquaforest Searchlight service on or off. The service is needed for Audit and OCR.

## 6.6 Searchlight.config file

The **Searchlight.config** file contains advanced settings that should only be updated from guidance of the support team ([support@aquaforest.com](mailto:support@aquaforest.com)). The file is located in the following location: “[installation path]\config\Searchlight.config”.

If a setting in the config file is updated, the Searchlight service must be restarted by going to **Settings > Advanced** and turning the service off and on again.

Some of the common settings available in the Searchlight.config file are described below.

Setting	Description
skipEnumerationErrors	Set this to <b>true</b> to skip documents that can't be enumerated due to permissions restrictions, long path errors, etc. instead of failing the whole job.
checkServiceEvery	This interval to periodically check the status of the Searchlight service. If the status of a job is set to as running when the service has stopped, it will be put into an error state. The default is to check the service every 60 minutes.
enumerationMaxParallelism	When enumerating documents from large SharePoint libraries, Aquaforest Searchlight partitions the retrieval so that the documents are retrieved in chunks. These chunks can be retrieved in parallel which can significantly speed up enumeration. This setting is used to control the maximum number of chunks that can be retrieved at once. Note, however, that the maximum value will be limited to the maximum cores your license permits.
deleteDocumentsAfterAudit	If the processing mode is “Audit and OCR” and there is enough space in the local computer where the Temp Folder is defined, 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.

Setting	Description
processSharepointList	By default, Searchlight only processes SharePoint document libraries. Set this setting to "true" if you want to process attachments in SharePoint Lists as well.
skipCheckedOutDocument	Set this to true to skip checked-out documents from being processed (during OCR stage only).
retainApprovalStatus	When Aquaforest Searchlight processes documents in a SharePoint library which requires Content Approval, it will set them to 'Pending' after processing. Set this setting to "true" to retain the original Approval Status after the documents have been processed.
ignorePreviouslyOcredDocuments	<p>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.</p> <p>To avoid re-processing these documents again irrespective of whether the modified has changed, set this setting to "true". The default value is false.</p>
sharePointFailCheckinComment	<p>When a SharePoint document is successfully OCR'd, a comment indicating the file was processed by Aquaforest Searchlight is added during check-in. This check-in comment can be configured in the "Library Settings" tab. However, when a document failed to OCR, no comment is added.</p> <p>To force Searchlight to add a comment to the non-OCR'd document in SharePoint, specify a comment in this setting.</p>

Setting	Description
failOnPixelLimit	<p>Force a document to error out in Native mode if it has an image in a page that exceeds the pixel limit (IRIS engine only). The default value is 'false' which will cause the page to be skipped.</p> <p>Extended OCR has the following image limits:</p> <ul style="list-style-type: none"> <li>• Max Height = 32,768 pixels</li> <li>• Max Width = 32,768 pixels</li> <li>• Max Size = 75,000,000 pixels</li> </ul>
pdfTextOperators	<p>The PDF text operators that need to be present in a page to consider it searchable.</p>
downloadAndUploadRetries sharePointRequestRetries	<p>Occasionally, there might be some intermittent network problems or unusual extreme load on the SharePoint server which can cause problems when processing SharePoint document libraries. To cope with this, retry mechanisms have been implemented for different scenarios that will retry performing a particular task in the event of such problems (e.g., timeouts). There are 2 SharePoint retry settings available:</p> <ul style="list-style-type: none"> <li>• downloadAndUploadRetries - used when downloading and uploading documents fail.</li> <li>• sharePointRequestRetries - used when executing SharePoint queries fail.</li> </ul> <p>The number of retries and the amount of time to wait between retries can be controlled through the respective config settings. The value needs to be entered in the format "x,y", where x is the number of retries and y is the time (in milliseconds) to wait before the first retry). For subsequent retries, the time to wait will be twice the previous wait time.</p>

Setting	Description
databaseRetries	<p>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 this setting.</p> <p>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.</p>

## 7 Acknowledgements

This product makes use of several Open-Source components which are included in binary form. The appropriate acknowledgements and copyright notices are given below.

Name	Homepage
AutoMapper	<a href="#">Homepage</a>   <a href="#">GitHub</a>
AvalonEdit	<a href="#">Homepage</a>   <a href="#">GitHub</a>
BitMiracle.LibTiff.NET	<a href="#">Homepage</a>   <a href="#">GitHub</a>
BouncyCastle.Crypto	<a href="#">Homepage</a>
ByteSize	<a href="#">GitHub</a>
Common.Logging	<a href="#">Homepage</a>
CompareNETObjects	<a href="#">GitHub</a>
CronExpressionDescriptor	<a href="#">Homepage</a>
Dapper	<a href="#">Homepage</a>   <a href="#">GitHub</a>
Extended.Wpf.Toolkit	<a href="#">Homepage</a>
IKVM.NET	<a href="#">Homepage</a>   <a href="#">Sourceforge</a>
Log4Net	<a href="#">Homepage</a>
MahApps MahApps.Metro MahApps.Metro.IconPacks	<a href="#">Homepage</a> <a href="#">GitHub</a> <a href="#">GitHub</a>
MailKit	<a href="#">GitHub</a>
MimeKit	<a href="#">GitHub</a>
Microsoft.WindowsAPICodePack.Core	<a href="#">Homepage</a>
Microsoft.WindowsAPICodePack.Shell	<a href="#">Homepage</a>
Modern UI (Metro) Charts	<a href="#">CodePlex</a>
Newtonsoft.Json	<a href="#">Homepage</a>
OpenMcdf	<a href="#">GitHub</a>
PDFBox	<a href="#">Homepage</a>
PnP-Sites-Core	<a href="#">GitHub</a>
Quartz	<a href="#">Homepage</a>   <a href="#">GitHub</a>
System.Data.SQLite	<a href="#">Homepage</a>
Tika	<a href="#">Homepage</a>
ZXing.Net	<a href="#">Homepage</a>