
Aquaforest Searchlight Reference Guide



Version 1.22
April 2017

Contents

1	Product Overview	3
1.1	The Business Problem: Documents that are not searchable.	3
1.2	The Solution: Aquaforest Searchlight	3
2	Installation and Licensing	4
2.1	System Requirements	4
2.2	SharePoint Online (Office 365) System Requirements	4
2.3	Licensing.....	4
2.3.1	Entering License Keys	5
3	Aquaforest Searchlight Modules.....	6
3.1	Multi-core Module	6
3.2	OCR Engines Modules	6
3.2.1	Aquaforest OCR Module (Included with the standard product).....	6
3.2.2	Extended (IRIS) OCR Module (Included with the standard product)	6
3.2.3	Extended OCR Asian Language Module (Extra Cost)	6
3.2.4	Extended OCR Advanced Compression (Extra Cost)	6
4	Searchlight Architecture and Concepts	7
4.1	Supported Formats	8
4.2	Searchlight Libraries.....	8
4.3	Searchability Status.....	8
4.4	Audit and Candidate Identification	9
4.5	Archiving	9
4.6	SharePoint and Office 365 Document Stores Concepts.....	9
4.6.1	Versioning	9
4.6.2	URL format	9
4.7	File System Document Stores Concepts.....	9
4.7.1	File Name Length.....	10
4.7.2	File Access Permissions.....	10
4.8	Aquaforest Searchlight Service	10
5	Quick Start Guide.....	11
5.1	Creating a Library	11
5.1.1	Library Settings	11
5.1.2	Document Settings.....	12
5.1.3	Document Archive Settings	13

5.1.4	OCR Settings.....	13
5.1.4.1	Extended OCR Engine Settings.....	13
5.1.4.2	Aquaforest OCR Engine Settings	14
5.1.5	Scheduler.....	14
5.1.6	Alert Settings.....	15
5.1.7	Finish	17
5.2	Updating a Library	18
5.3	Audit & Conversion Status	19
6	The Aquaforest Searchlight Tool	21
6.1	Welcome Screen	21
6.2	Dashboard.....	22
6.3	Library	23
6.3.1	Library Status	23
6.3.2	Library Settings	24
6.3.3	Document Settings.....	25
6.3.3.1	Retain Creation/Modified Date/User.....	27
6.3.3.2	SharePoint Libraries (Retain Creation/Modified Date/User)	30
6.3.4	Document Archive Settings	31
6.3.5	OCR Settings.....	31
6.3.5.1	Aquaforest OCR Settings	32
6.3.5.2	Extended OCR Settings.....	34
6.3.6	Run Details	38
6.3.7	Run Details Context Menu.....	38
6.3.8	Scheduler Settings	39
6.3.9	Alert Settings.....	40
6.4	Help & Support	41
6.4.1	Diagnostic Tool.....	42
6.5	Settings	42
6.5.1	License Settings.....	42
6.5.2	Email Settings.....	43
6.5.3	Themes	44
6.5.4	Advanced Settings.....	44
6.6	Searchlight.config file.....	45
7	Acknowledgements	48

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 doesn't 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

Supported Operating Systems	<ul style="list-style-type: none">• Windows 7 (x64)• Windows 8 (x64)• Windows 10(x64)• Windows Server 2008 R2 (x64)• Windows Server 2012 R2 (x64)• Windows Server 2016
Supported Document Stores	<ul style="list-style-type: none">• SharePoint 2010• SharePoint 2013• SharePoint 2016• SharePoint Online (Office 365)• Windows File Systems
Disk Space	950 MB
Memory	Minimum 4GB (recommended 8GB)
Visual C++ Redistributable	Visual C++ 2010 Redistributable (x86 x64) and Visual C++ 2012 Redistributable (x86 x64)
.NET Framework	3.5 and 4.5.2

2.2 SharePoint Online (Office 365) System Requirements

Supported Operating Systems	Windows 7 SP1 and above (x64) Windows 8 (x64) Windows 10 (x64) Windows Server 2008 R2 SP1 and above (x64) Windows Server 2012 (x64) Windows Server 2016
Additional tools	SharePoint Server Client Components SDK (x86 x64)

2.3 Licensing

Aquaforest Searchlight has 3 main licensing levels:

- Single Core
- 8 Cores
- 32 Cores

Further Modules are also available upon request. These are:

- Multi-core module with more than 32 cores (up to a limit of 64)
- Intelligent High Quality Compression
- Asian Languages OCR support

Trial licenses usually are time limited, that is, it will expire after a particular 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.

You either don't have a license or your current license is invalid

Please contact support@aquaforest.com to request a new license. If you already have a new license, enter it in the text box below and click Ok.

Email 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.

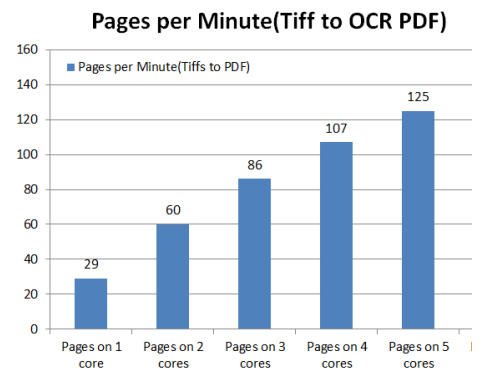
3 Aquaforest Searchlight Modules

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 two OCR Engines namely the Aquaforest OCR Engine and the Extended (IRIS) OCR Engine. Below is an explanation of the OCR Engines.

3.2.1 Aquaforest OCR Module (Included with the standard product)

The Aquaforest 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 Aquaforest 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 Language Support
- Optional IHQC Advanced PDF Compression



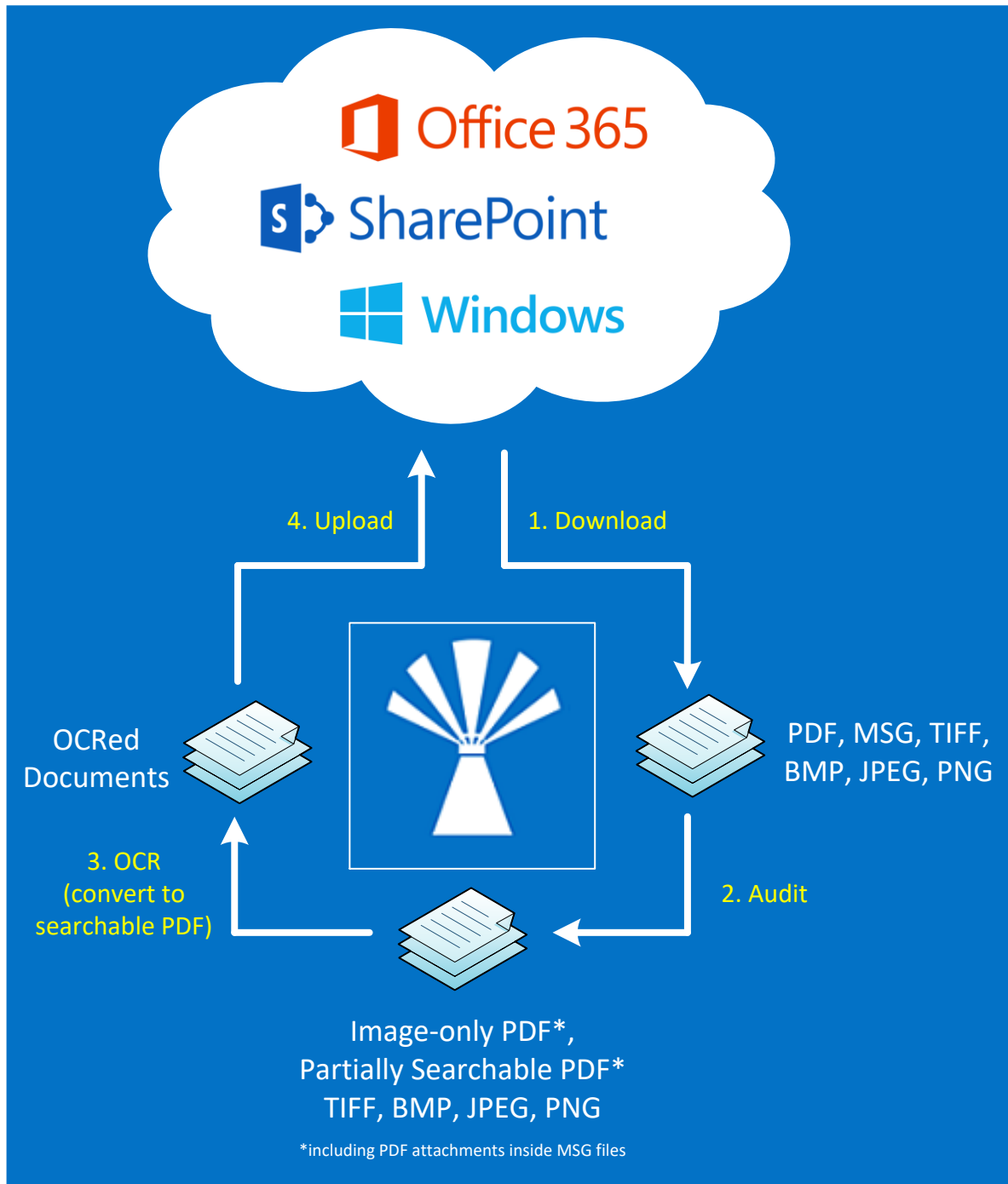
3.2.3 Extended OCR Asian Language Module (Extra Cost)

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

3.2.4 Extended OCR Advanced Compression (Extra Cost)

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. download (for SharePoint locations) or copy (for file system locations) documents from the source location to a temp location,
2. analyse (Audit) them to identify whether they need to be OCREd,
3. OCR them to convert them to searchable PDFs and
4. put them back in the source location, replacing the existing documents

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

4.1 Supported Formats

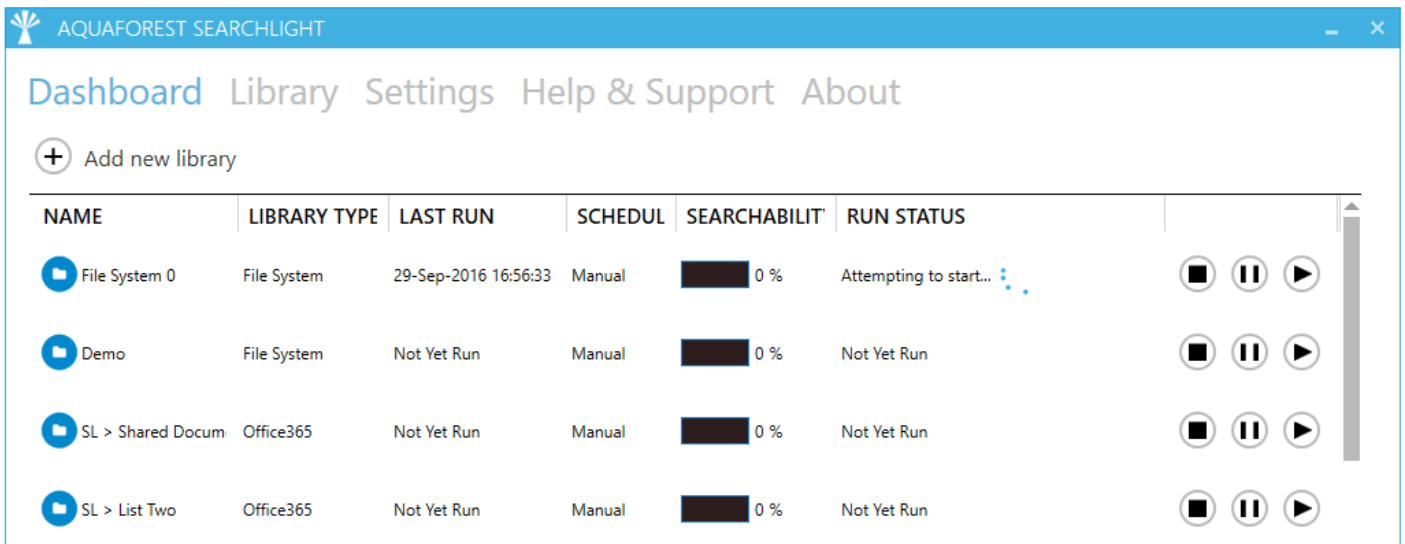
Aquaforest Searchlight currently supports TIFF, BMP, JPG, PNG and PDF documents (including PDF attachments inside MSG files) as input. As a result, candidate documents will always be one of these formats.

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. At the top, there is a navigation bar with the following links: Dashboard (active), Library, Settings, Help & Support, and About. Below the navigation bar, there is a button labeled '+ Add new library'. The main content area contains a table with the following columns: NAME, LIBRARY TYPE, LAST RUN, SCHEDULE, SEARCHABILITY, and RUN STATUS. The table lists four libraries:

NAME	LIBRARY TYPE	LAST RUN	SCHEDULE	SEARCHABILITY	RUN STATUS	
File System 0	File System	29-Sep-2016 16:56:33	Manual	0 %	Attempting to start...	Stop, Pause, Play
Demo	File System	Not Yet Run	Manual	0 %	Not Yet Run	Stop, Pause, Play
SL > Shared Docum	Office365	Not Yet Run	Manual	0 %	Not Yet Run	Stop, Pause, Play
SL > List Two	Office365	Not Yet Run	Manual	0 %	Not Yet Run	Stop, Pause, Play

A Searchlight library should not be confused with a SharePoint document library, which is a document library in SharePoint.

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:

- Fully Searchable
A PDF document is fully searchable if all its pages have text that can be indexed and searched
- Partially Searchable
A partially searchable document contains some pages with text, others with only images or no images and no text (blank)
- Image-only
This is a PDF that has been created from one or more images – most commonly as a result 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 in order 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 Archiving

To avoid making inadvertent changes to the source document, it is recommended to turn Archiving on so as to have a backup of the source documents. Archiving is the process of copying over the source documents to an archive location specified by the user before performing any sort of processing on them.

4.6 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.

4.6.1 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.

4.6.2 URL format

Below is an example of how to set the SharePoint URL format when setting up a document library in Searchlight.

Actual URL	Valid URL
http(s)://SharePoint2010/site/myLibrary/myForms/AllItems.aspx	http(s)://SharePoint2010/site/myLibrary
http(s)://SharePoint2013/site/Library/_layouts/15/start.aspx#/mylibrary/Forms/AllItems.aspx	http(s)://SharePoint2013/site/Library/mylibrary

4.7 File System Document Stores Concepts

Aquaforest Searchlight can be configured to monitor folders on the windows file system. Below are a few issues that need to be considered when using the Windows File System.

4.7.1 File Name Length

The windows operating system has a limit to file name length it can process. Aquaforest Searchlight always runs an audit before any conversion is carried out. Before the start of an audit, if any files with long names are found it will be reported to the user and the user can either shorten the file names or move the files.

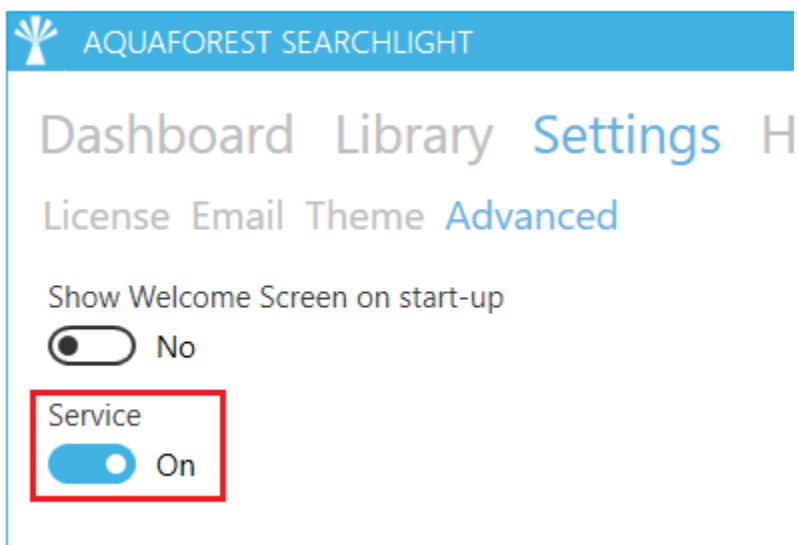
4.7.2 File Access Permissions

If there are any secured locations that are to be monitored, you will have to configure the Aquaforest Searchlight Service with the security credentials of a user that has permissions to access that particular location.

4.8 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 OCR'd. 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.



5 Quick Start Guide

5.1 Creating a Library

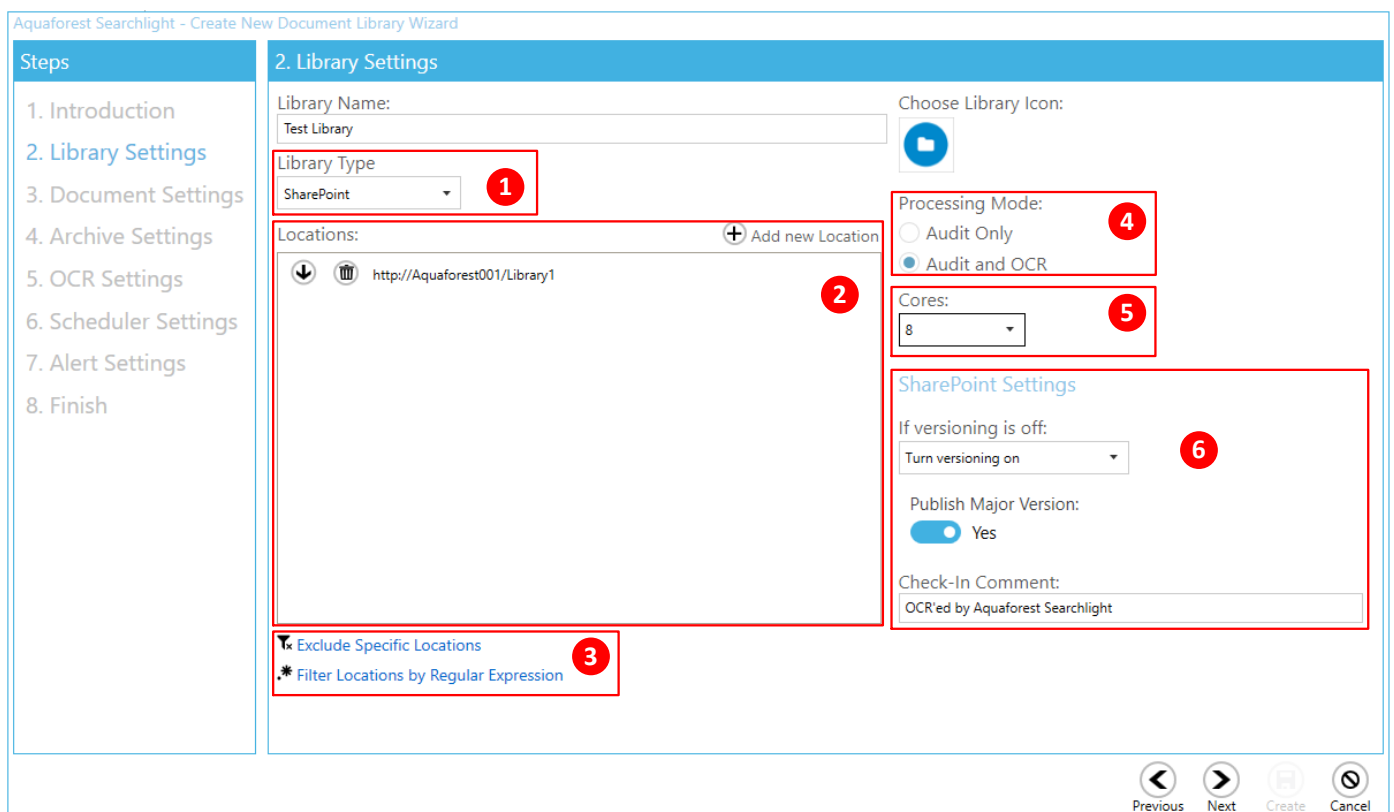
Creating Document Libraries in Aquaforest Searchlight is managed by a wizard. This wizard can be launched by clicking the **Add new library** button on the Dashboard.

Dashboard Library Settings



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



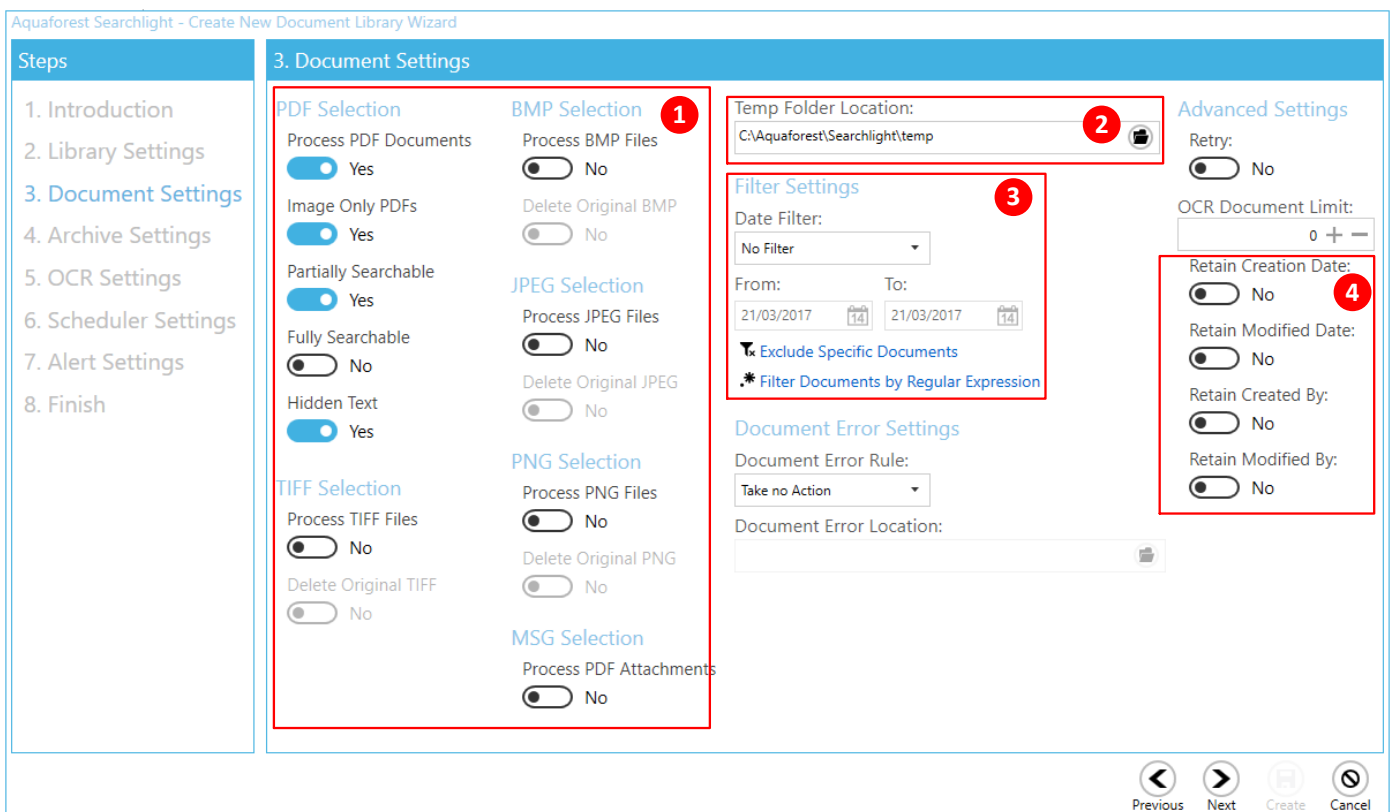
1. Do the documents to be processed reside on a file system, SharePoint or Office 365?
2. Enter the location(s). It can consist of the following:
 - o one or more SharePoint site collections, SharePoint sites, SharePoint document libraries and/or SharePoint lists.
 - OR
 - o one or more File System paths
3. There are 2 ways to filter locations:
 - a. Excluding specific locations – locations that match the specified URL(s) are excluded
 - b. By regular expressions – locations (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.

- Do you only want to **Audit Only**, or **Audit and OCR**? Audit means that Searchlight will analyze 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.
- The number of cores to use to process documents in parallel. For instance, if 8 cores is specified, Searchlight will process 8 documents simultaneously, which will significantly reduce the total processing time.
- Turn versioning on if you want to have a 'backup' of the original documents, otherwise the documents will be overwritten with new searchable ones.

5.1.2 Document Settings

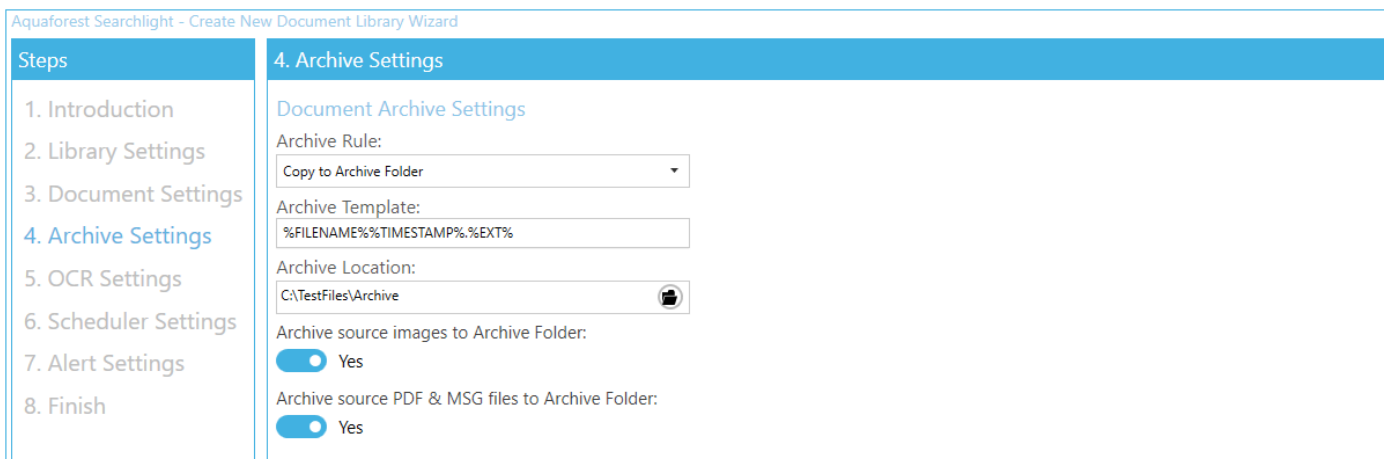
This page enables the user to specify rules and criteria for the selection of documents to be processed.



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

5.1.3 Document Archive Settings

This page provides the option of archiving source files so as to have a backup before OCR is applied to them.



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

4. Archive Settings

Document Archive Settings

Archive Rule:
Copy to Archive Folder

Archive Template:
%FILENAME%%TIMESTAMP%.%EXT%

Archive Location:
C:\TestFiles\Archive

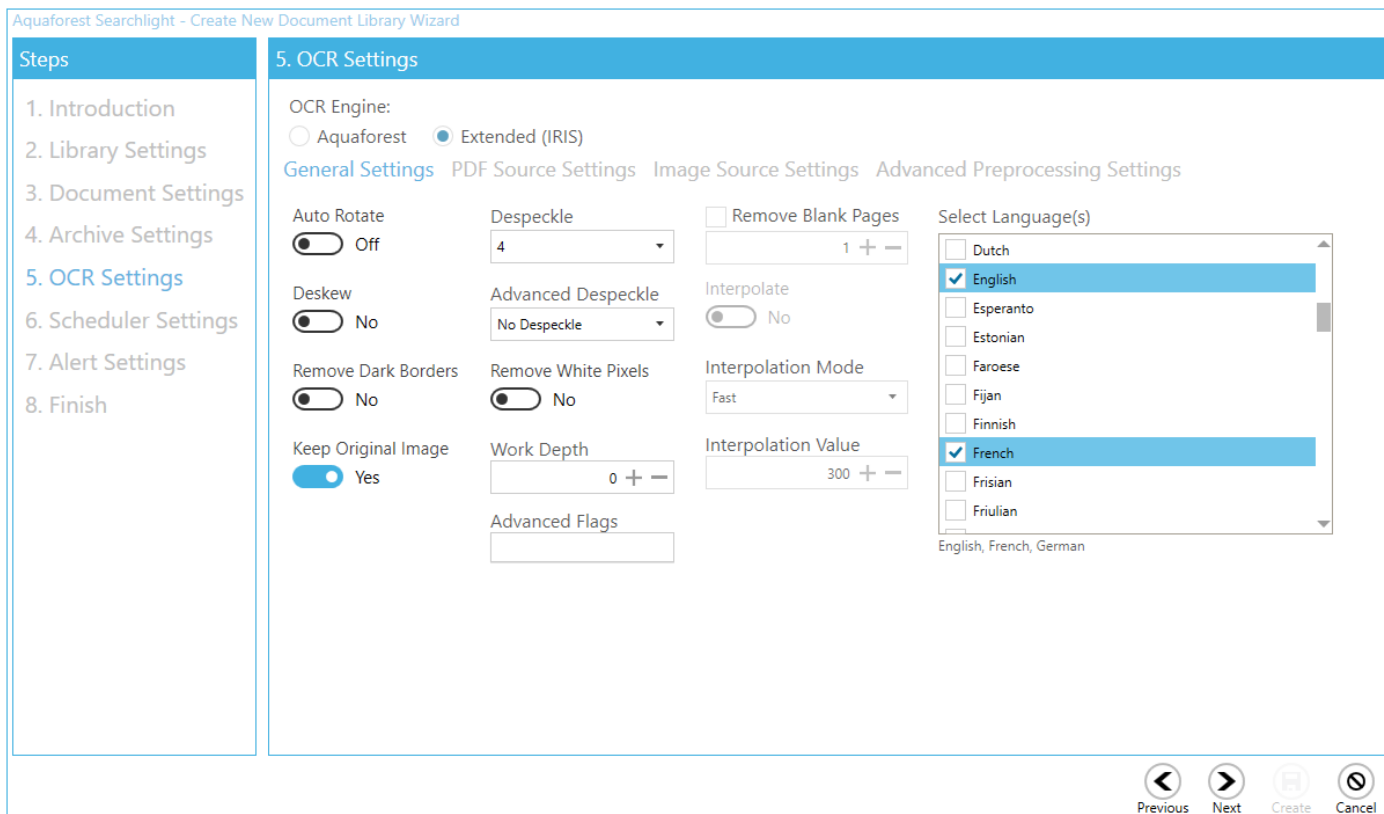
Archive source images to Archive Folder:
 Yes

Archive source PDF & MSG files to Archive Folder:
 Yes

5.1.4 OCR Settings

In this section, you can set the OCR settings. Aquaforest Searchlight comes bundled with two OCR Engines: Aquaforest OCR engine and the Extended IRIS (Canon) OCR engine. The Extended OCR is the default engine and supports more languages (100+) than the Aquaforest OCR engine and also has the ability to 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

Interpolate: No

Interpolation Mode: Fast

Interpolation Value: 300

Deskew: No

Advanced Despeckle: No Despeckle

Remove Dark Borders: No

Remove White Pixels: No

Work Depth: 0

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 Aquaforest 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

Deskew On OCR Language

Remove Lines Off Box Graphics

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 particular time or every hour etc.

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

6. Scheduler Settings

Manual
 Once per day
 Continuous
 Run once

At:

Every:

Between And

On:

At:

Previous Next Create 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 particular date range. Creating alerts is managed by another wizard within the library creation wizard.

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

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

What action(s) do you want the alert task to perform?

Send an email Yes

Generate a CSV report Yes

Attach the CSV report to the email Yes

Save Report No

Location:

Previous Next

Previous Next Create Cancel

2. Select the email 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

7. Alert Settings

Configuration

Action > Email

Email Settings

From Email Address:

To Email Address:

Email Subject:

Email Message: Test Email

Further processing of '%LIBRARYNAME%' has been suspended.
Log file: %LOGFILEPATH%

Previous Next

Previous Next Create Cancel

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

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

Action > Report

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

Show details of individual documents that were processed

No

Choose the columns that will appear in the report.

Previous Next

Previous Next Create Cancel

- Select when you want the task to run. Based on the current settings, you will get an email with the report attached 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

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

Daily
 Weekly
 Monthly
 One time

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

Day(s):

The: Last Friday

Advanced Settings

On Job Success No

On Job Error No

Expires

Previous Next

Previous Next Create Cancel

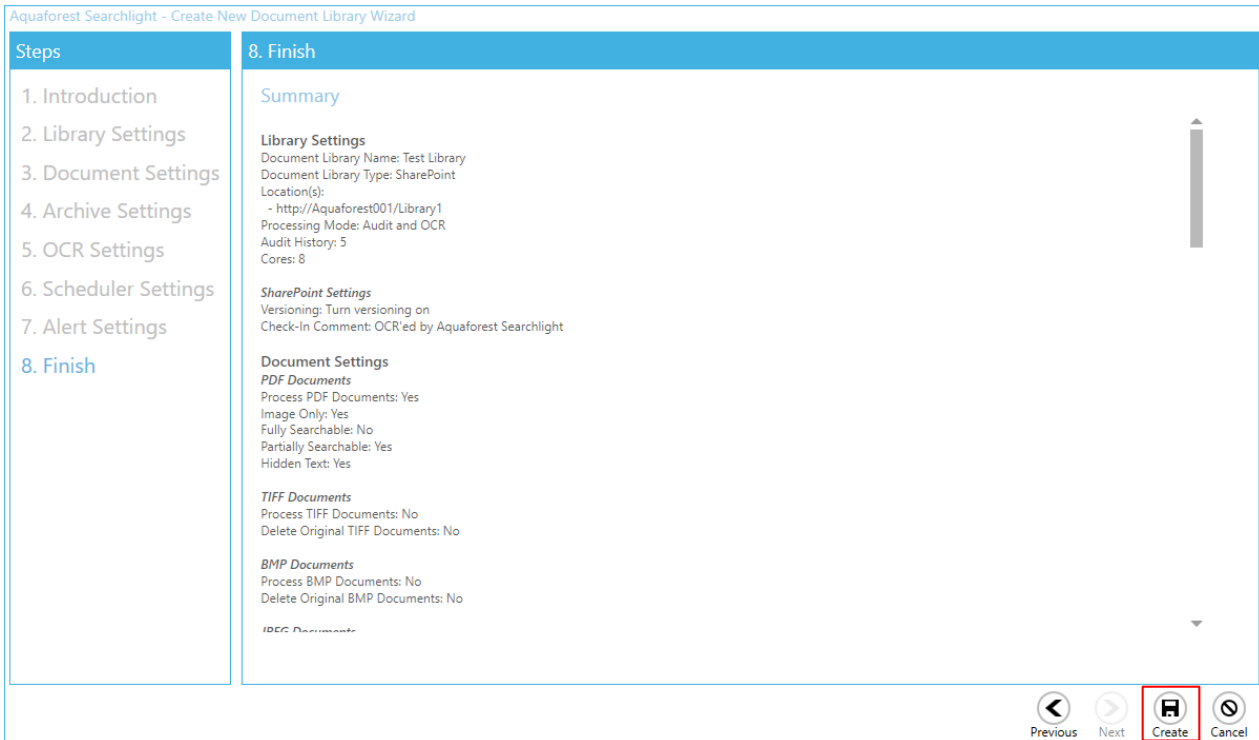
Aquaforest Searchlight 1.22 Reference Guide

Page 16

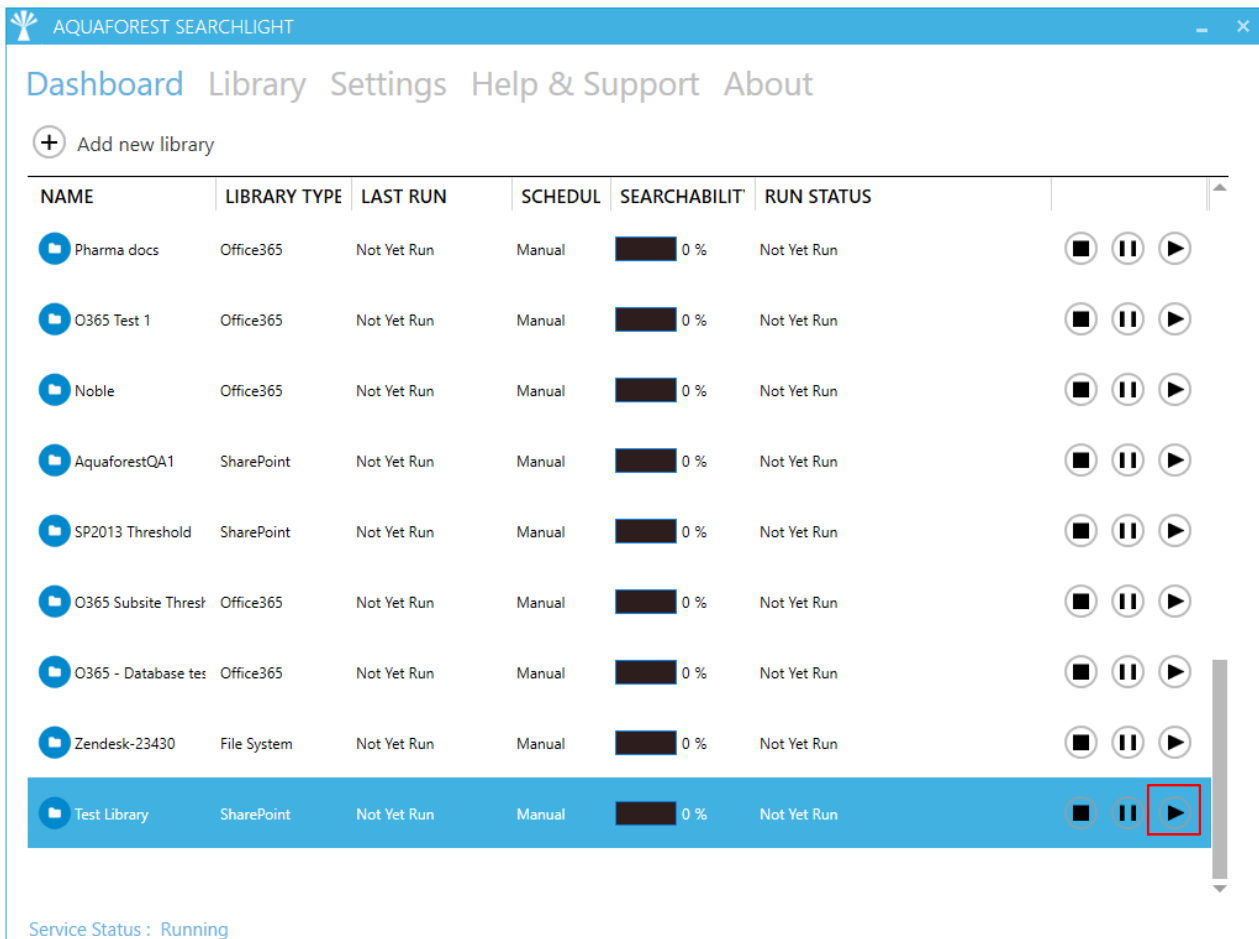
Aquaforest

5.1.7 Finish

In 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.



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

1. All the settings of a library can be edited by double-clicking the library from the dashboard.
2. You can also select a library to edit by choosing the library from the combo box at the top of the page.
3. 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 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 File System

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

STATISTICS

PDF Documents

Total PDF Documents:	80
Image-only PDFs:	6 (7.6 %)
Partially Searchable PDFs:	12 (15.2 %)
Fully Searchable PDFs:	61 (77.2 %)
Error PDF Documents:	1
Total PDF Pages:	13,265
Image-only Pages:	307 (2.3 %)
Fully Searchable Pages:	12,958 (97.7 %)

Image (TIFF,BMP,JPG,PNG) Documents

MSG Documents

Library Totals

Total Documents:	80
Total Error Documents:	1
Total Pages:	13,265
Total Searchable Pages:	12,958 (97.7 %)

Generate Audit Report

LOG OUTPUT

number of image pages: 1, number of searchable pages 1787.

Checking document: C:\Aquaforest\OCRSDK\samples\documents\output\PDF\Set2\MA-CICLO1_200808_T1_2_searchable.pdf
Searchability Status : searchable, Number of Pages 5044.

Document library statistics after audit:

PDF Documents

Total PDF Documents:	80
Image-only PDFs:	6 (7.6 %)
Partially Searchable PDFs:	12 (15.2 %)
Fully Searchable PDFs:	61 (77.2 %)
Error PDF Documents:	1
Total PDF Pages:	13,265
Image-only Pages:	307 (2.3 %)
Fully Searchable Pages:	12,958 (97.7 %)

Image (TIFF, BMP, JPEG and/or PNG) Documents

Total Image Documents:	0
Error Image Documents:	0
Total Image Pages:	0

Library Totals

Total Documents:	80
Total Error Documents:	1
Total Pages:	13,265
Total Searchable Pages:	12,958 (97.7 %)

24-Aug-2016 11:02:22: Audit ended

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 shows the Aquaforest Searchlight web 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:

- Top right: "Showing last 5 runs" (Callout 1).
- Table with columns: #, RUN ID, RUN DATE, PROCESSING MODE, Status, AUDIT RESULTS (Successful Documents, Error Documents), Status, CONVERSION RESULTS (Successful Documents, Error Documents).
- Row 1: #1, RUN ID 31014, RUN DATE 09-Nov-2015 12:33:44, PROCESSING MODE Audit and OCR, Status Completed, Successful Documents 156687, Error Documents 65, Status Completed, Successful Documents 78470, Error Documents 20.

Run Details Section:

- Top right: Radio buttons for "Audit" (selected) and "Conversion" (Callout 2).
- Table with columns: #, DOCUMENT PATH, SEARCHABILITY, STATUS, MODIFIED, PAGES.
- Row 94: #94, DOCUMENT PATH http://win-cujcfc16f2q/sites/searchlight/QA%202/qa/221112/split/chk_images002945.tif, SEARCHABILITY imageonly, STATUS, MODIFIED Nov-2015 14:20:01, PAGES 1.
- Row 95: #95, DOCUMENT PATH http://win-cujcfc16f2q/sites/searchlight/QA%202/qa/221112/split/chk_images002953.tif, SEARCHABILITY imageonly, STATUS, MODIFIED Nov-2015 14:19:33, PAGES 1.
- Row 96: #96, DOCUMENT PATH http://win-cujcfc16f2q/sites/searchlight/QA%202/qa/221112/split/chk_images002921.tif, SEARCHABILITY imageonly, STATUS, MODIFIED Nov-2015 14:20:13, PAGES 1.
- Row 97: #97, DOCUMENT PATH http://win-cujcfc16f2q/sites/searchlight/QA%202/qa/221112/split/chk_images002919.tif, SEARCHABILITY imageonly, STATUS, MODIFIED Nov-2015 14:19:54, PAGES 1.
- Row 98: #98, DOCUMENT PATH http://win-cujcfc16f2q/sites/searchlight/QA%202/qa/221112/split/chk_images002924.tif, SEARCHABILITY imageonly, STATUS, MODIFIED Nov-2015 14:20:15, PAGES 1.
- Row 99: #99, DOCUMENT PATH http://win-cujcfc16f2q/sites/searchlight/QA%202/qa/221112/split/chk_images002947.tif, SEARCHABILITY imageonly, STATUS, MODIFIED Nov-2015 14:19:43, PAGES 1.
- Row 100: #100, DOCUMENT PATH http://win-cujcfc16f2q/sites/searchlight/QA%202/qa/221112/split/chk_images002952.tif, SEARCHABILITY imageonly, STATUS, MODIFIED Nov-2015 14:19:45, PAGES 1.
- Row 101: #101, DOCUMENT PATH http://win-cujcfc16f2q/sites/searchlight/QA%202/qa/221112/split/chk_images002923.tif, SEARCHABILITY imageonly, STATUS, MODIFIED Nov-2015 14:19:54, PAGES 1.
- Row 102: #102, DOCUMENT PATH http://win-cujcfc16f2q/sites/searchlight/QA%202/qa/221112/split/chk_images002918.tif, SEARCHABILITY imageonly, STATUS, MODIFIED Nov-2015 14:20:03, PAGES 1.
- Bottom right: "Limit 500" (Callout 4).
- Bottom right: "Export to CSV", "Generate Log", "View Full Log", "Reload" buttons (Callout 6).
- Bottom left: Navigation buttons for "1" (Callout 5).
- Callout 3: A searchability filter dropdown menu is open, showing options: Searchable, Image Only, Partially Searchable, Error, and Hidden Text. "Partially Searchable" is selected.

Service Status : 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 particular 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 provide assistance in creating the first document library.

Welcome to Aquaforest Searchlight Version 1.20




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.

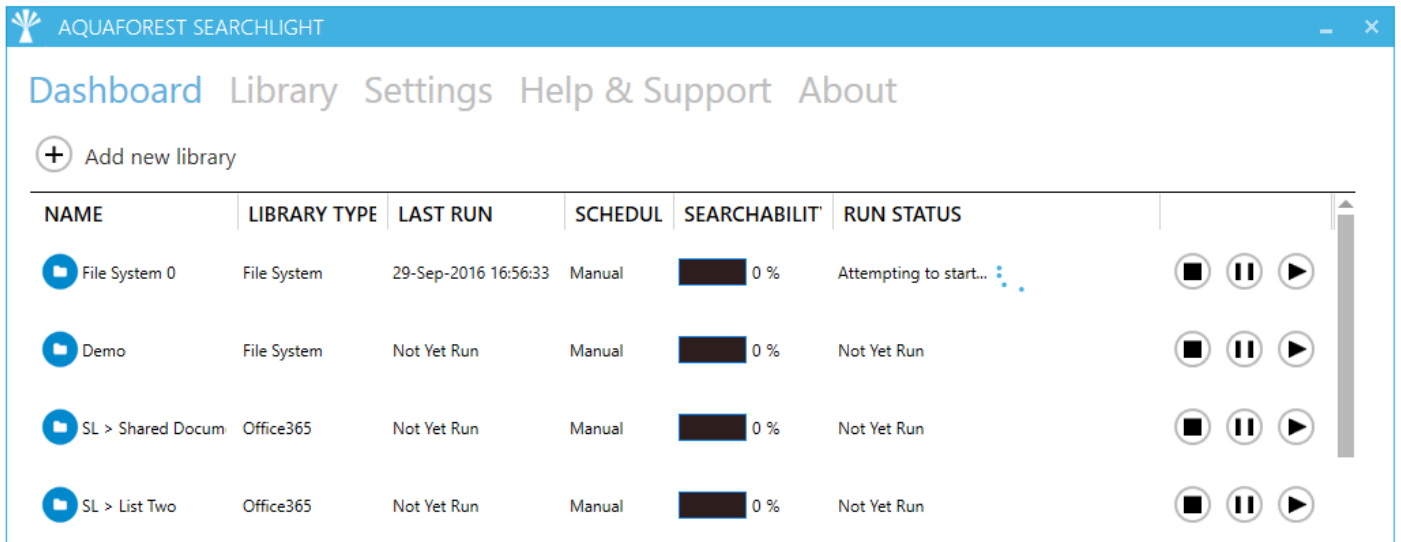
There is a sample Searchlight Document Library which you can process to get an understanding of how the product works and the [Reference Guide](#) provides more detailed information.


© Aquaforest Limited 2001-2016

Show this message on startup Yes

[Continue](#)

6.2 Dashboard



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"> • SharePoint • Office 365 • File System
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
Status	Current status of the document library: <ul style="list-style-type: none"> • Running • Completed • Error • Aborted
■ ▶	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.

AQUAFOREST SEARCHLIGHT

Dashboard **Library** Settings Help & Support About

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

STATISTICS

PDF Documents

Total PDF Documents:	80
Image-only PDFs:	6 (7.6 %)
Partially Searchable PDFs:	12 (15.2 %)
Fully Searchable PDFs:	61 (77.2 %)
Error PDF Documents:	1
Total PDF Pages:	13,265
Image-only Pages:	307 (2.3 %)
Fully Searchable Pages:	12,958 (97.7 %)

Image (TIFF,BMP,JPG,PNG) Documents

MSG Documents

Library Totals

Total Documents:	80
Total Error Documents:	1
Total Pages:	13,265
Total Searchable Pages:	12,958 (97.7 %)

Generate Audit Report

LOG OUTPUT

number of image Pages: 1, number of Searchable Pages 1787.

Checking document: C:\Aquaforest\OCRSDK\samples\documents\output\PDF\Set2\MA-CICLO1_200808_T1_2_searchable.pdf
Searchability Status : searchable, Number of Pages 5044.

Document library statistics after audit:

PDF Documents
Total PDF Documents: 80
Image-only PDFs: 6 (7.6 %)
Partially Searchable PDFs: 12 (15.2 %)
Fully Searchable PDFs: 61 (77.2 %)
Error PDF Documents: 1
Total PDF Pages: 13,265
Image-only Pages: 307 (2.3 %)
Fully Searchable Pages: 12,958 (97.7 %)

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

Library Totals
Total Documents: 80
Total Error Documents: 1
Total Pages: 13,265
Total Searchable Pages: 12,958 (97.7 %)

24-Aug-2016 11:02:22: Audit ended

Service Status : Running

6.3.2 Library Settings

AQUAFOREST SEARCHLIGHT
-
x

Dashboard **Library Settings** Settings Help & Support About
O365 - Library 5

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

Library Name:

Library Type

Locations: + Add new Location

↑ 🗑️ https://aquafores[REDACTED]

User Name
Password

Choose Library Icon:

Processing Mode:
 Audit Only
 Audit and OCR

Cores:

SharePoint Settings

If versioning is off:

Publish Major Version:
 Yes

Check-In Comment:

🗑️ Exclude Specific Locations

* Filter Locations by Regular Expression

🗑️ Delete
🔄 Refresh
💾 Save

Service Status : Stopped

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
Locations	One or more locations 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.

Setting	Description
Processing Mode	<ul style="list-style-type: none"> Audit Only Analyse the document library to find out the documents that need to be converted without actually converting them Audit & OCR Perform audit on the document library and OCR the documents that have been identified as candidates for processing
Cores	This determines the maximum number of CPU cores that will be used when running the job.
SharePoint Versioning	This setting can be used to automatically turn versioning on.
Publish Major Version	Publish major version after OCR
Check-in Comment	The check-in comment applied to the updated SharePoint file version.

6.3.3 Document Settings

The screenshot displays the 'Document Settings' page in the Aquaforest Searchlight application. The page is organized into several sections:

- PDF Selection:** Includes toggle switches for 'Process PDF Documents' (Yes), 'Image Only PDFs' (Yes), 'Partially Searchable' (Yes), 'Fully Searchable' (No), and 'Hidden Text' (Yes).
- TIFF Selection:** Includes toggle switches for 'Process TIFF Files' (Yes) and 'Delete Original TIFF' (No).
- BMP Selection:** Includes toggle switches for 'Process BMP Files' (No) and 'Delete Original BMP' (No).
- JPEG Selection:** Includes toggle switches for 'Process JPEG Files' (No) and 'Delete Original JPEG' (No).
- PNG Selection:** Includes toggle switches for 'Process PNG Files' (No) and 'Delete Original PNG' (No).
- MSG Selection:** Includes a toggle switch for 'Process PDF Attachments' (No).
- Temp Folder Location:** A text input field containing 'C:\Aquaforest\Searchlight\temp'.
- Filter Settings:** Includes a 'Date Filter' dropdown set to 'No Filter', and date pickers for 'From' and 'To' (both set to 23/01/2017).
- Advanced Settings:** Includes toggle switches for 'Retry' (No), 'Retain Creation Date' (No), 'Retain Modified Date' (No), 'Retain Created By' (No), and 'Retain Modified By' (No). It also features an 'OCR Document Limit' input field set to 0.
- Document Error Settings:** Includes a 'Document Error Rule' dropdown set to 'Take no Action' and a 'Document Error Location' text input field.

At the bottom left, the 'Service Status' is indicated as 'Stopped'. At the bottom right, there are 'Refresh' and 'Save' buttons.

Setting	Description
Process PDF	Whether or not to process PDF documents

Setting	Description
Image Only	Whether or not to process Image-only PDFs. 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.
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	Whether or not process PDF documents with hidden text in them. 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. Note: If you set this setting to true, you might want to consider setting Remove Hidden Text to true in the "OCR Settings > PDF Source Settings", otherwise you will have multiple OCR text layers per page.
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	The operation to perform if a document fails to process: <ul style="list-style-type: none"> • Copy to error folder • Move to error folder (for file system library type only)
Document Error Location	The path of the error location

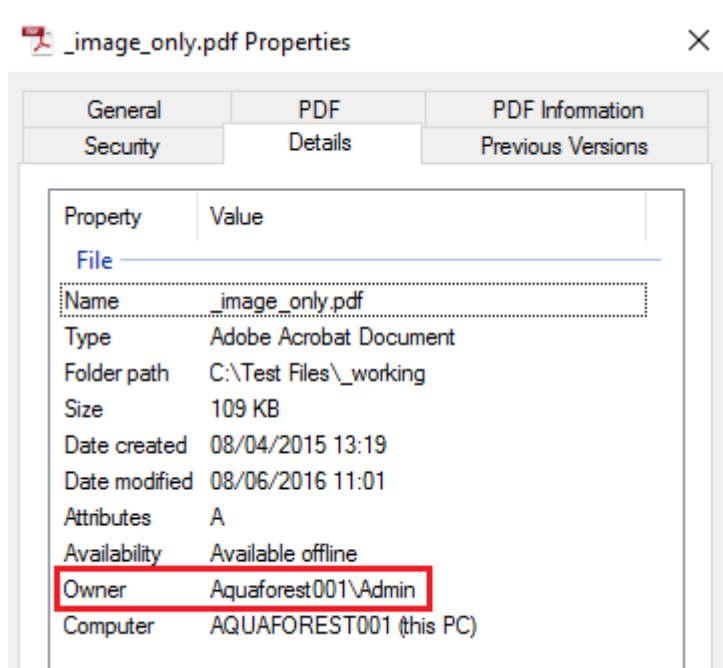
Setting	Description
Retry	Whether or not to re-process documents that have previously failed to convert
OCR Document Limit	Limit the number of documents to OCR (not Audit) per run. Set to '0' for no limits.
Retain Creation Date*	Retain the creation date of the source document (SharePoint creation date, 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](#) and [6.3.3.2](#) 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 would need to be running with sufficient administrative privileges.

- **** SharePoint metadata vs. PDF metadata**

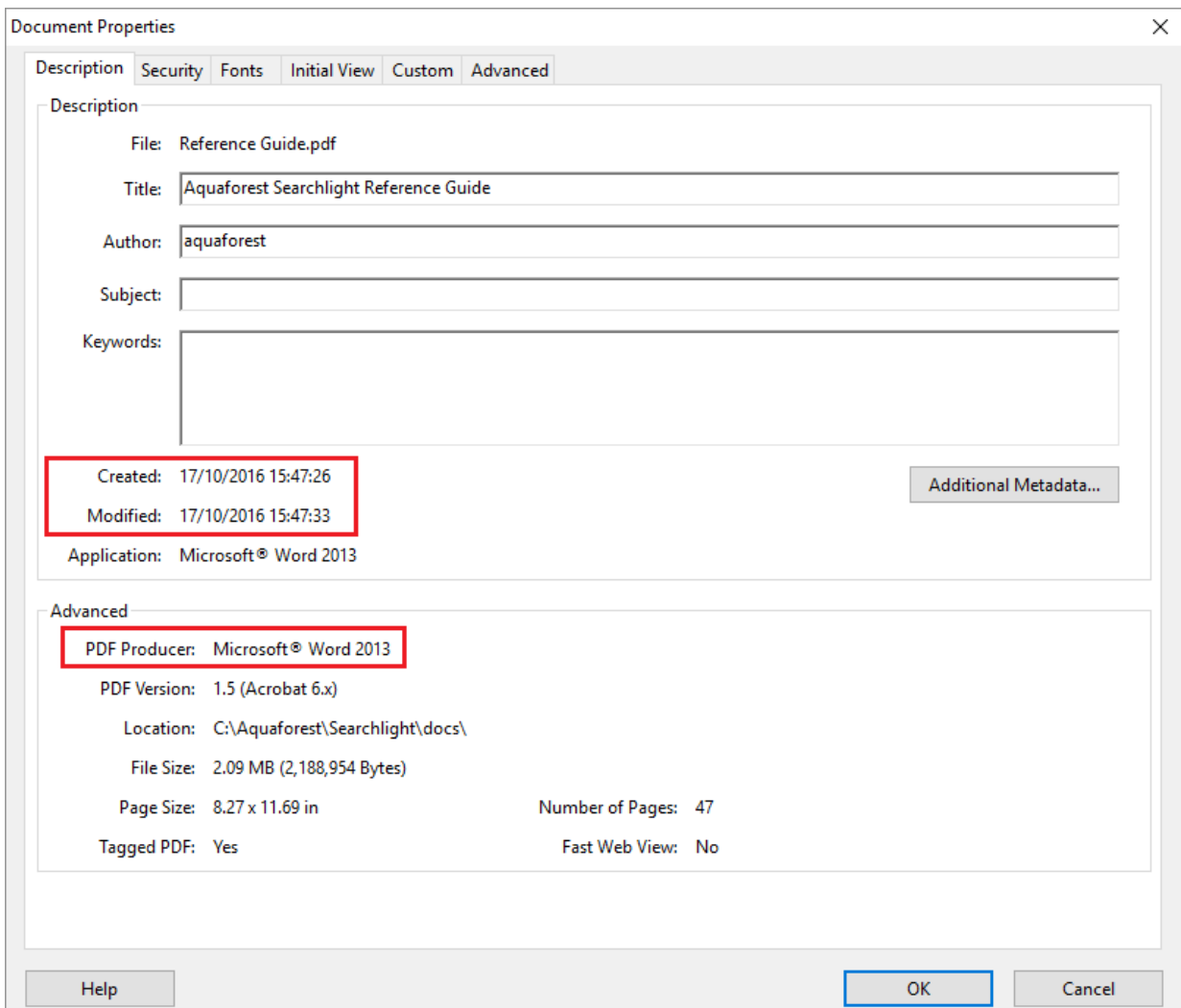
SharePoint metadata refers to the 'columns' available in SharePoint that stores information about each document.

Columns

A column stores information about each document in the document library. The following columns are currently available in this document library:

Column (click to edit)	Type
Title	Single line of text
IM Address	Single line of text
Modified	Date and Time
Created	Date and Time
Created By	Person or Group
Modified By	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 (Retain Creation/Modified Date/User)

The behaviour of Retain Creation/Modified date and Retain Approval Status (Searchlight.config) can vary depending on the settings used in SharePoint and Searchlight. The table below summarises when these will and will not be retained.

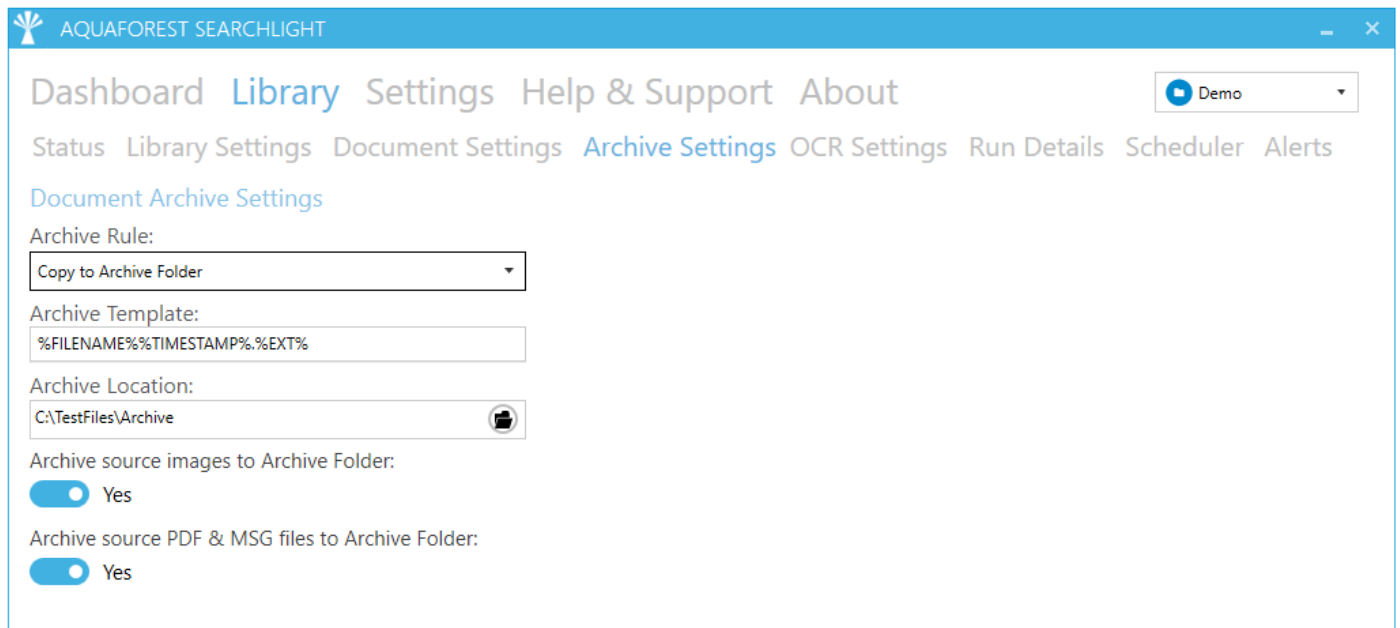
SharePoint Settings				Searchlight Settings		Created Date retained?	Created By retained?	Modified Date retained?	Modified By retained?
Create Major Versions	Create Minor Versions	Require Content Approval	Require Checkout	Retain Approval Status	Publish Major Version				
x	x	x	x	n/a**	n/a*	✓	✓	✓	✓
x	x	x	✓	n/a**	n/a*	✓	✓	✓	✓
x	x	✓	x	x	n/a*	✓	✓	✓	✓
				✓	n/a*	✓	✓	✓	✓
x	x	✓	✓	x	n/a*	✓	✓	✓	✓
				✓	n/a*	✓	✓	x	✓
✓	x	x	x	n/a**	n/a*	✓	✓	✓	✓
✓	x	x	✓	n/a**	n/a*	✓	✓	✓	✓
✓	x	✓	x	x	n/a*	✓	✓	✓	✓
				✓	n/a*	✓	✓	x	✓
✓	x	✓	✓	x	n/a*	✓	✓	✓	✓
				✓	n/a*	✓	✓	x	✓
✓	✓	x	x	n/a**	x	✓	✓	✓	✓
				n/a**	✓	✓	✓	✓	x
✓	✓	x	✓	n/a**	x	✓	✓	✓	✓
				n/a**	✓	✓	✓	✓	x
✓	✓	✓	x	x	x	✓	✓	✓	✓
				x	✓	✓	✓	✓	x
				✓	x	✓	x	✓	x/✓*
				✓	✓	✓	✓	✓	x/✓*
✓	✓	✓	✓	x	x	✓	✓	✓	✓
				x	✓	✓	✓	✓	x
				✓	x	✓	x	✓	x/✓*
				✓	✓	✓	✓	✓	x/✓*

n/a*	To publish major version, both major and minor versioning must be on in SharePoint
n/a**	To retain moderation status, 'Require Content Approval' must be on in SharePoint
x/✓*	"Modified By" will be retained if the original moderation status of the document being OCR'd was set to "Approved". If it was set to "Draft" or "Pending", "Modified By" won't be retained.
	When both "Retain Approval Status" and "Publish Major Version" are enabled in Searchlight, "Retain Approval Status" precedes "Publish Major Version". E.g.: Usually if the original "Approval Status" is set to "Draft" in SharePoint and major version is published, the "Approval Status" will change to "Pending". However, in this case (both "Retain Approval Status" and "Publish Major Version" is set to true in Searchlight), major version will not be published and the original "Approval Status" (Draft) will be retained. If the original "Approval Status" is set to "Pending" or "Approved", then retaining the approval status automatically publishes the document. If the original "Approval Status" is set to "Pending" then the user will still have to manually approve or reject the status.

In summary:

- Retaining approval status (Searchlight setting) when minor versioning is on (SharePoint setting) will NOT retain "Modified Date".
- Publishing major version will NOT retain "Modified By".
- If versioning is off or only major versioning is on in SharePoint, 'Modified By' will be retained but if BOTH major AND minor versioning is on in SharePoint, 'Modified By' will NOT be retained (with exception – see "x/✓*" in table above)

6.3.4 Document Archive Settings



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.

6.3.5 OCR Settings

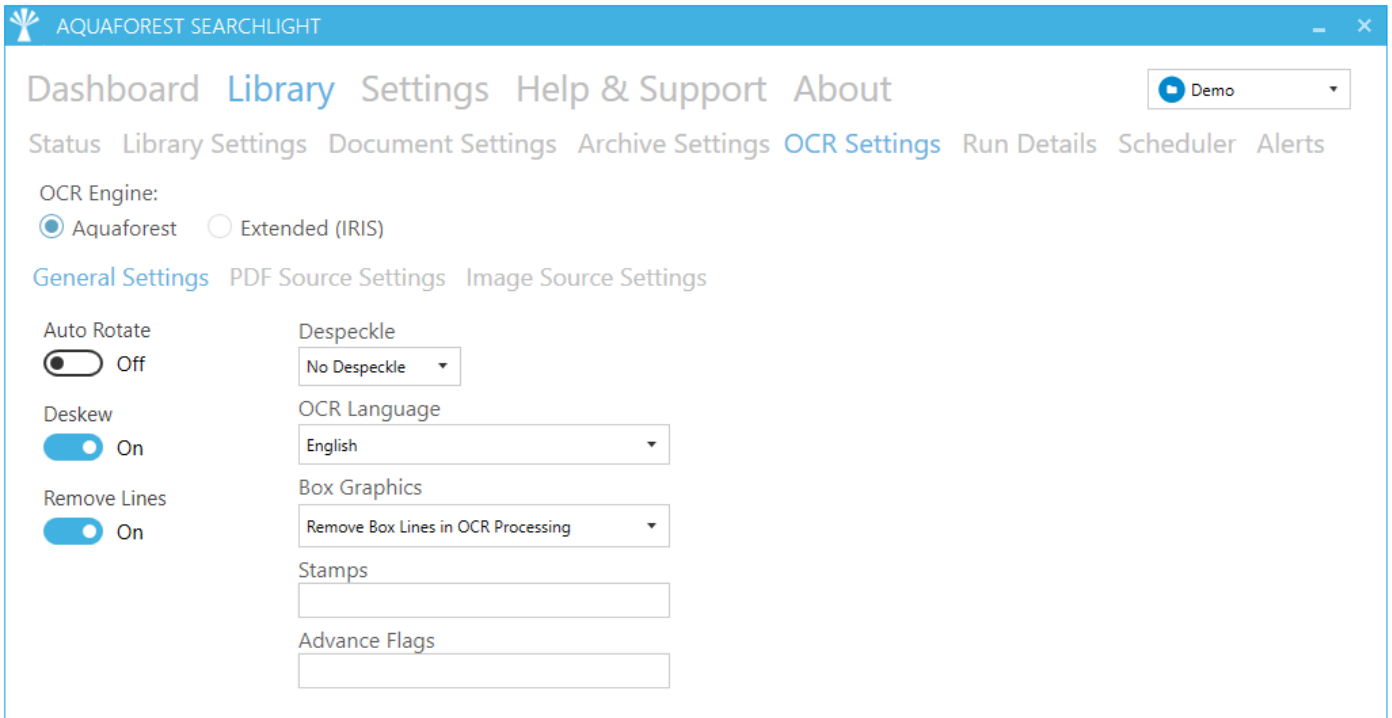
As described in [section 5.1.4](#), Aquaforest Searchlight has 2 OCR engines. When creating a new library, the default OCR settings are loaded from the Properties.xml file for each OCR engine.

- 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 Aquaforest OCR Settings



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

Setting	Description														
Stamps	<p>The "Stamps" parameter allows entry of a command-line style specification.</p> <p>For example, the string below will produce a stamp "Page Number000123Final" on page 1, "Page Number000124Final" on page 2, etc. Note the need to use escaped quotes for prefixes and suffixes with spaces.</p> <pre>/stamppref="\Page Number\" /stampsuff=Final /stampstart=123 /stampdigits=6 /stamppos=0 /stamptype=0</pre> <table border="1" data-bbox="608 539 1490 1532"> <thead> <tr> <th>Parameter</th> <th>Description</th> </tr> </thead> <tbody> <tr> <td>/stamppref</td> <td>Prefix – a string to be added to the beginning of the stamp, before the number section.</td> </tr> <tr> <td>/stampsuff</td> <td>Suffix - a string to be added to the end of the stamp, after the number section.</td> </tr> <tr> <td>/stampstart</td> <td>Start – the value that the number portion of the stamp should start at. The number portion will be incremented by 1 each page.</td> </tr> <tr> <td>/stampdigits</td> <td>Digits – a value indicating the minimum length that the number portion of the stamp should be displayed as. Preceding 0's will be used to pad any numbers less than this whilst numbers greater than this will be displayed in full.</td> </tr> <tr> <td>/stamppos</td> <td>Stamp Position : <ul style="list-style-type: none"> • 0 is TopLeft, • 1 is TopCenter, • 2 is TopRight, • 3 is CenterLeft, • 4 is Center, • 5 is CenterRight, • 6 is BottomLeft, • 7 is BottomCenter, • 8 is BottomRight </td> </tr> <tr> <td>/stamptype</td> <td>Stamp Type : <ul style="list-style-type: none"> • 0 stamp is added as text • 1 stamp is added as an image </td> </tr> </tbody> </table>	Parameter	Description	/stamppref	Prefix – a string to be added to the beginning of the stamp, before the number section.	/stampsuff	Suffix - a string to be added to the end of the stamp, after the number section.	/stampstart	Start – the value that the number portion of the stamp should start at. The number portion will be incremented by 1 each page.	/stampdigits	Digits – a value indicating the minimum length that the number portion of the stamp should be displayed as. Preceding 0's will be used to pad any numbers less than this whilst numbers greater than this will be displayed in full.	/stamppos	Stamp Position : <ul style="list-style-type: none"> • 0 is TopLeft, • 1 is TopCenter, • 2 is TopRight, • 3 is CenterLeft, • 4 is Center, • 5 is CenterRight, • 6 is BottomLeft, • 7 is BottomCenter, • 8 is BottomRight 	/stamptype	Stamp Type : <ul style="list-style-type: none"> • 0 stamp is added as text • 1 stamp is added as an image
Parameter	Description														
/stamppref	Prefix – a string to be added to the beginning of the stamp, before the number section.														
/stampsuff	Suffix - a string to be added to the end of the stamp, after the number section.														
/stampstart	Start – the value that the number portion of the stamp should start at. The number portion will be incremented by 1 each page.														
/stampdigits	Digits – a value indicating the minimum length that the number portion of the stamp should be displayed as. Preceding 0's will be used to pad any numbers less than this whilst numbers greater than this will be displayed in full.														
/stamppos	Stamp Position : <ul style="list-style-type: none"> • 0 is TopLeft, • 1 is TopCenter, • 2 is TopRight, • 3 is CenterLeft, • 4 is Center, • 5 is CenterRight, • 6 is BottomLeft, • 7 is BottomCenter, • 8 is BottomRight 														
/stamptype	Stamp Type : <ul style="list-style-type: none"> • 0 stamp is added as text • 1 stamp is added as an image 														
Advanced Flags	Command line flags to be passed through to the underlying executable. Contact support@aquaforest.com for details on using this field.														
PDF Source Settings															
Re-Image PDF	Each page of the source PDF is rasterized to an image and appended to a new PDF document.														
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'.														

Setting	Description
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.
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.
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.
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

AQUAFORREST SEARCHLIGHT

Dashboard Library Settings Help & Support About Demo

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

OCR Engine:
 Aquaforest Extended (IRIS)

General Settings PDF Source Settings Image Source Settings Advanced Preprocessing Settings

Auto Rotate: Off
 Deskew: Yes
 Remove Dark Borders: No
 Keep Original Image: Yes

Despeckle: No Despeckle
 Advanced Despeckle: No Despeckle
 Remove White Pixels: No
 Work Depth: 0 + -
 Advanced Flags:

Remove Blank Pages: 1 + -
 Interpolate: No
 Interpolation Mode: Normal
 Interpolation Value: + -

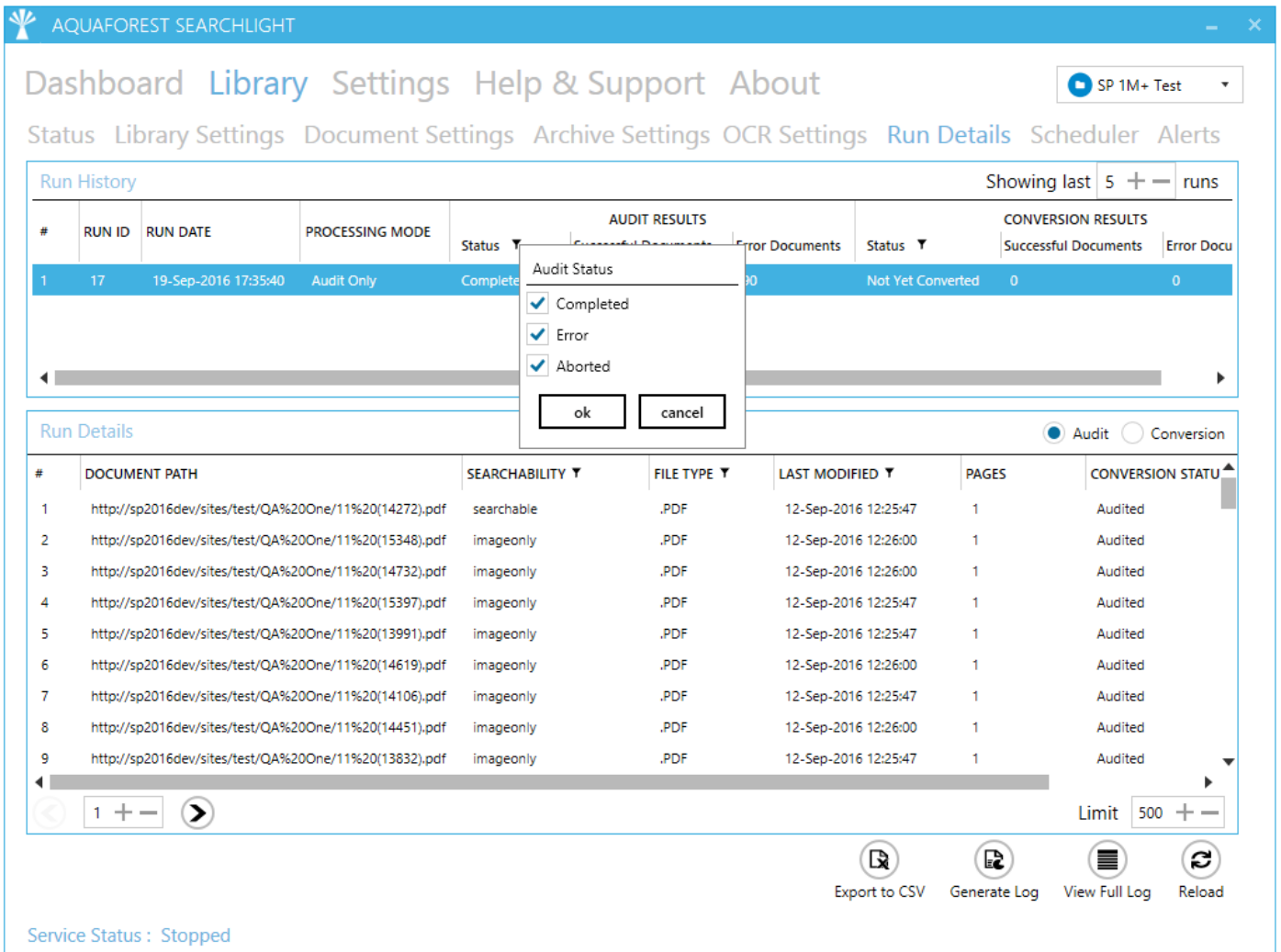
Select Language(s)
 Danish
 Dutch
 English
 Esperanto
 Estonian
 Faroese
 Fijian
 Finnish
 French
 Frisian
 English, French

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. Note: 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 pre-processing is applied. Note: 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.
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, less blank pages are detected.
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.
Language	Set the language(s) to use for OCR. Note: <ul style="list-style-type: none"> • Only a maximum of 8 languages can be selected • Only the English language can be used in conjunction with an Asian language
PDF Source Settings	
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'.

Setting	Description
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.
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.
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.
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.

Setting	Description
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).

6.3.6 Run Details



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.

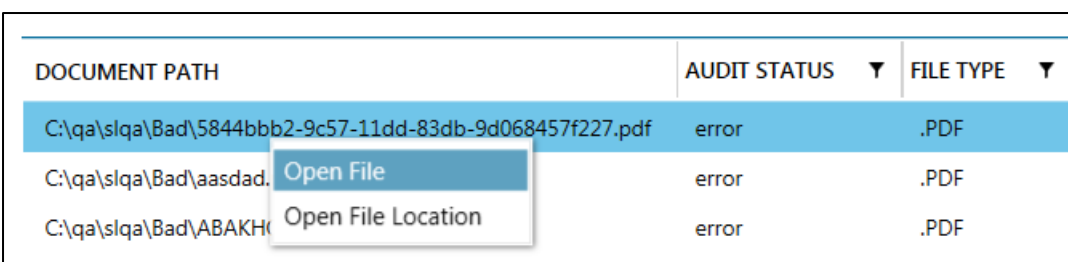
Use **Export to CSV** to export the run details to CSV file.

The **Generate Log** button is used to generate a log report of the selected run as a PDF, RTF or HTML file.

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

6.3.7 Run Details Context Menu

Using the Right-Click context menu on a particular file allows the file or file location to be opened.



6.3.8 Scheduler Settings

AQUAFORREST SEARCHLIGHT

Dashboard Library Settings Help & Support About

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

Manual

Once per day

At: 16 : 50

Continuous

Every: 7 Days

Between 00 : 01 And 23 : 59

Run once

On: 12/06/2014

At: 16 : 50

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

6.3.9 Alert Settings

Configuration

- Action
- Email
- Report
- Trigger
- Finish

Action

What action(s) do you want the alert task to perform?

Send an email
 Yes

Generate a CSV report
 Yes

Attach the CSV report to the email
 Yes

Save Report
 No

Location:

Service Status : Running

View Alert Log Refresh Save

Setting	Description
Action	
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
Save Report	Save the report locally
Location	The location to save the report
Email	
From Email Address	The email address to send the email from.
To Email Address	The email address to send the email to.
Email Subject	The email subject. You can use the following templates: %LIBRARYNAME% - will be replaced by the name of the library %STATUS% - will be replaced by "success" or "error" depending on whether the job ran successfully or not

Setting	Description
Email Message	The email message to send. You can use the following templates: %LIBRARYNAME% - will be replaced by the name of the library %STATUS% - will be replaced by "success" or "error" depending on whether the job ran successfully or not %LOGFILEPATH% - will be replaced by the path of the log file.
Trigger	
On Job Success	Run the alert task every time the library runs successfully
On Job Error	Run the alert task every time the library fails to run successfully
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

Reference Guide
The [reference guide](#) contains detailed information about the product.

Support
For product technical support, send us an email at support@aquaforest.com or call us on +44 (0)1296 768 727.

Release Notes
See the [release notes](#) to see the changes made in the different versions of Aquaforest Searchlight.

Sales
For sales and pricing matters, send us an email at sales@aquaforest.com or call us on +44 (0)1296 768 727.

Troubleshooting Guide
The [troubleshooting guide](#) contains common configuration issues that can affect the operation of Aquaforest Searchlight.

Remote Session
Request a [remote session](#) if you want help setting up Aquaforest Searchlight on your system.

Blogs
The Aquaforest Searchlight [blogs](#) contain tips and best practices to get the best out of the product.

Live Chat
You can always contact us on [live chat](#) during office hours.

Diagnostics Tool
Run the [diagnostics tool](#) to see if your system meets all the requirements to run Aquaforest Searchlight successfully.

Estimate OCR Time
You can check this [blog](#) or [email](#) us details about the types of documents you wish to process, no. of pages and available hardware and we'll provide you with an estimate.

Service Status : Running

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

In order to run the diagnostic tool, click on the “Diagnostics Tool” icon in the “Help & Support” tab as pointed out in the image below. 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 particular document library. All the gathered information will be made available in a zip file which can be sent to support@aquaforest.com for further investigation.

The screenshot shows a grid of support options:

- Configuration Issues:** Icon of a wrench and screwdriver. Text: "configuration issues that can affect the operation of Aquaforest Searchlight."
- Blogs:** Icon of a document with a speech bubble. Text: "The Aquaforest Searchlight blogs contain advices on how to get the best of the product."
- Diagnostics Tool:** Icon of a gear. Text: "Run the [diagnostics tool](#) to see if your system meets all the requirements to run Aquaforest Searchlight successfully." (This option is highlighted with a red border in the original image).
- Live Chat:** Icon of two speech bubbles. Text: "You can always contact us on [live chat](#) during office hours."
- Estimate OCR Time:** Icon of a clock with a circular arrow. Text: "You can check this [blog](#) or [provide us](#) with the types of documents you wish to process, no. of pages and available hardware and we'll provide you with an estimate."

At the bottom left, it says "Service Status : Running".

6.5 Settings

6.5.1 License Settings

The screenshot shows the "License" settings page in Aquaforest Searchlight. The page title is "AQUAFOREST SEARCHLIGHT". The navigation menu includes "Dashboard", "Library", "Settings", "Help & Support", and "About". The "License" tab is selected, with sub-tabs for "Email", "Theme", and "Advanced".

License details:

- License Type: Permanent
- Computer Bound: X
- Multi-core: ✓
- Max Cores: 64
- Document Limit: Unlimited
- Trial Stamp: X
- Expires: X
- Features: Aquaforest OCR: ✓ | Extended OCR: ✓ (Asian Languages: ✓; IHQC: ✓)
- License Key: [Redacted] [Update]

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 OCRed. If this limit is reached, OCR will be disabled.

Setting	Description
Trial Stamp	Whether or not the OCR'd documents will have a trial stamp
Features	Modules enabled by the current license
License Key	The license key currently being used

6.5.2 Email Settings

The settings screen allows email server information to be defined. This is used to support the "Alerts" functionality.

AQUAFOREST SEARCHLIGHT

Dashboard Library **Settings** Help & Support About

License **Email** Theme Advanced

SMTP Server

SMTP Port

Username

Password

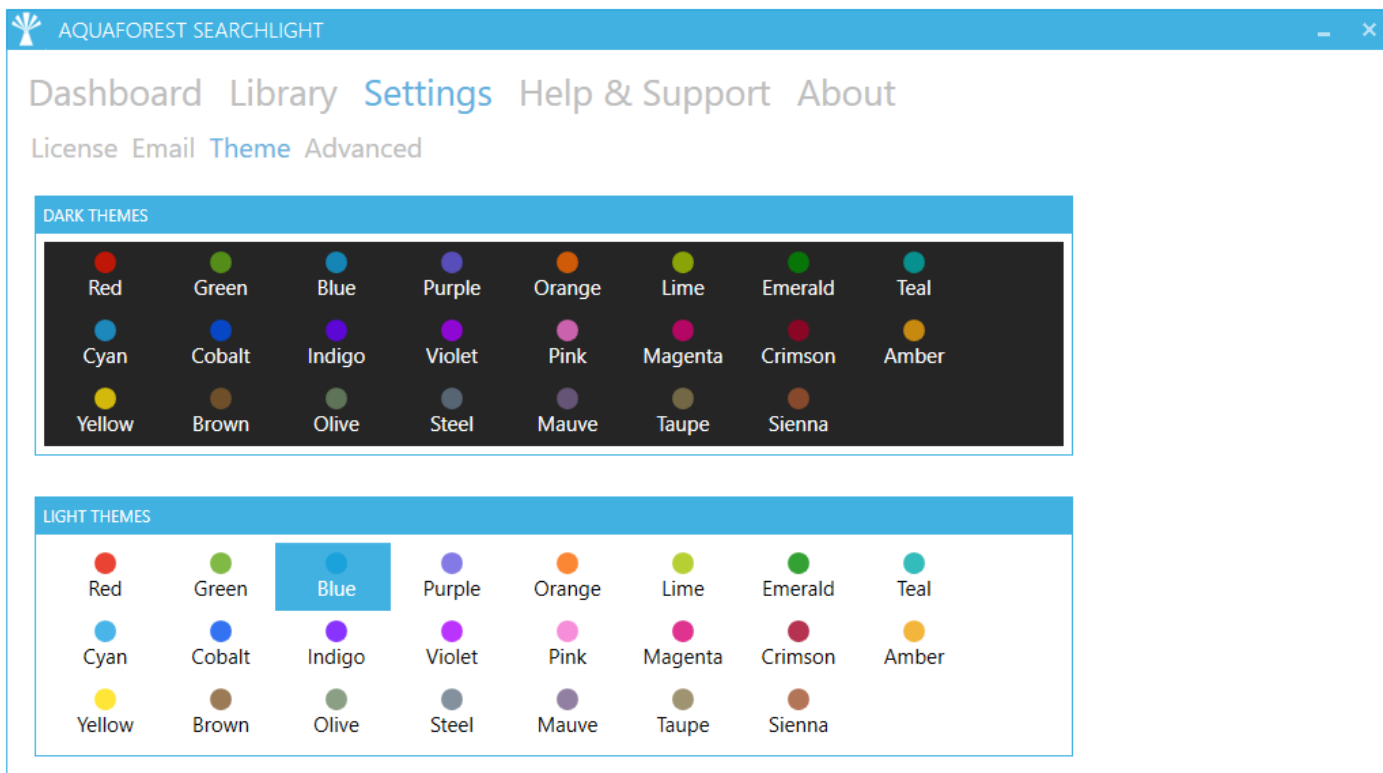
Re-enter Password

Refresh Save

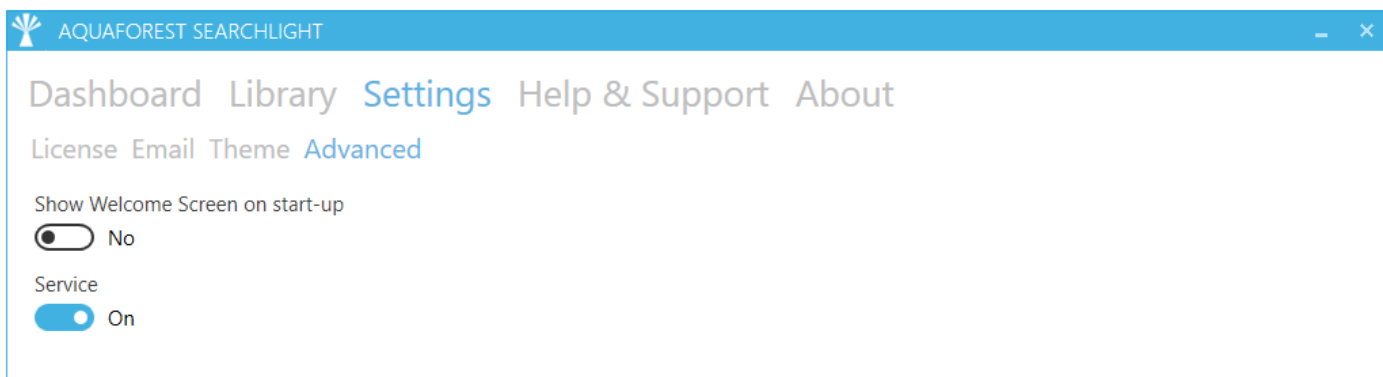
Setting	Description
SMTP Server	Address of the server hosting the SMTP server.
SMTP Port	SMTP Server port.
Username	Username for authentication by the server.
Password	Password for the username.

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 Advanced Settings



Setting	Description
Show Welcome Screen on start-up	Whether or not to show the Welcome Screen 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). 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 true 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.
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).

Setting	Description
retainApprovalStatus	<p>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.</p>
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>
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"> • Max Height = 32,768 pixels • Max Width = 32,768 pixels • Max Size = 75,000,000 pixels
pdfTextOperators	<p>The PDF text operators that need to be present in a page to consider it searchable.</p>

Setting	Description
<p>downloadAndUploadRetries sharePointRequestRetries</p>	<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"> • downloadAndUploadRetries - used when downloading and uploading documents fail • sharePointRequestRetries - used when executing SharePoint queries fail <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>
<p>databaseRetries</p>	<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 a number of Open Source components which are included in binary form. The appropriate acknowledgements and copyright notices are given below.

Name	Homepage
AvalonEdit	Homepage GitHub
BitMiracle.LibTiff.NET	Homepage GitHub
Cuneiform	n/a (Copyright (c) 1993-2008, Cognitive Technologies)
Common.Logging	Homepage
CompareNETObjects	GitHub
CronExpressionDescriptor	Homepage
DbLinq	GitHub CodePlex
Extended.Wpf.Toolkit	Homepage
FreedImage.NET	Homepage
IKVM.NET	Homepage Sourceforge
iTextSharp 4.1.6	Github
Leptonica	Homepage
Libjpeg	Homepage
Libpng	Homepage
Libtiff	Homepage
log4Net	Homepage
MahApps MahApps.Metro MahApps.Metro.IconPacks	Homepage GitHub GitHub
Microsoft.WindowsAPICodePack.Core	Homepage
Microsoft.WindowsAPICodePack.Shell	Homepage
Modern UI (Metro) Charts	CodePlex
OpenMcdf	Homepage
PDFBox	Homepage
Quartz	Homepage GitHub
RTF Writer	Homepage
System.Data.SQLite	Homepage
Zlib	Homepage