Aquaforest

Searchlight
Reference Guide

Aquaforest

Making Content Findable

Searchlight

Reference Guide

Version 2.2

November 2021

© Aquaforest Ltd 2001 - 2021

Content

1 PRODUCT OVERVIEW 5

1.1 The Business Problem: Documents that are not searchable. 5

1.2 The Solution: Aquaforest Searchlight 5

2 INSTALLATION AND LICENSING 6

2.1 System Requirements 6

2.2 SharePoint Online (Office 365) System Requirements 6

2.3 Licensing 6
2.3.1 Entering License Keys 7
2.3.2 Service Configuration 7

3 AQUAFOREST SEARCHLIGHT MODULES 9

3.1 Multi-core Module 9

3.2 OCR Engines Modules 9
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 (Extra Cost) 9
3.2.4 Extended Arabic & Farsi Languages Module (Extra Cost) 9
3.2.5 Extended Hebrew Language Support (Extra cost) 9
3.2.6 Extended OCR Advanced Compression (Extra Cost) 9

4 SEARCHLIGHT ARCHITECTURE AND CONCEPTS 10

4.1 Supported Formats 11

4.2 Searchlight Libraries 11

4.3 Searchability Status 11

4.4 Audit and Candidate Identification 12

4.5 Document Stores Concepts 12
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

4.6 Archiving **15**

4.7 Aquaforest Searchlight Service **15**

5 QUICK START GUIDE 16

5.1 Creating a Library **16**
5.1.1 Library Settings 17
5.1.2 Document Settings 18
5.1.3 Document Archive Settings 19
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

5.2 Updating a Library **25**

5.3 Importing settings from an existing Library **26**

5.4 Audit & Conversion Status **27**

6 THE AQUAFOREST SEARCHLIGHT TOOL 31

6.1 Welcome Screen **31**

6.2 Dashboard **32**

6.3 Library **33**
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

6.4 Help & Support **58**
6.4.1 Diagnostic Tool 59
6.4.2 Database Clean-up Tool 59

6.5 Settings **61**
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

6.6 Searchlight.config file **65**

7 ACKNOWLEDGEMENTS 69

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 store
- 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	• Windows 10(x64) • Windows Server 2012 R2 (x64) • Windows Server 2016 • Windows Server 2019
Supported Document Stores	• SharePoint 2010 • SharePoint 2013 • SharePoint 2016 • SharePoint 2019 • SharePoint Online (Office 365) • OneDrive for Business • Azure File Storage • Azure Blob Storage • Windows File Systems
Disk Space	950 MB
Memory	Minimum 4GB (recommended 8GB)
Visual C++ Redistributable	Visual C++ 2017 Redistributable (x86|x64)
.NET Framework	4.7.2

2.2 SharePoint Online (Office 365) System Requirements

Supported Operating Systems	Windows 10 (x64) Windows Server 2012 (x64) Windows Server 2016 Windows Server 2019
Additional tools	SharePoint Server Client Components SDK (x86|x64)

2.3 Licensing

Aquaforest Searchlight has 3 main licensing levels:

- Single Core
- 4 Cores
- 8 Cores

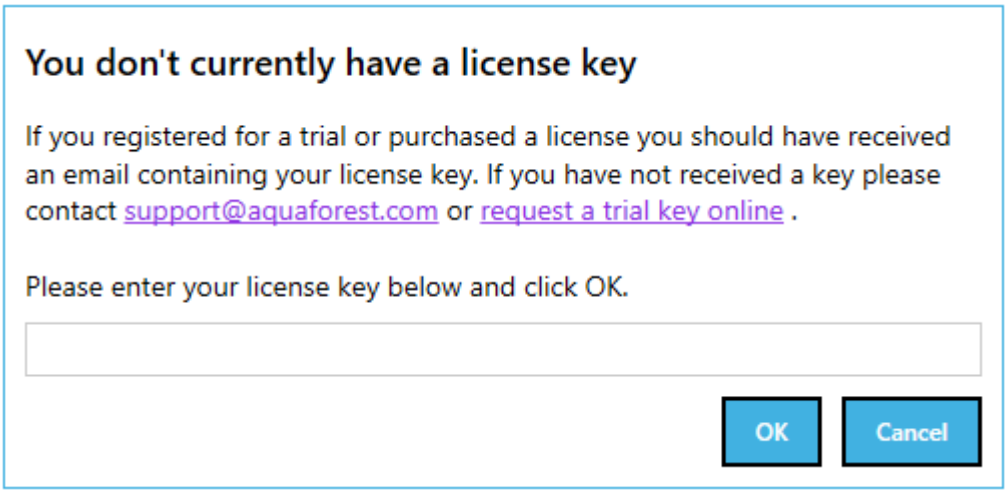
Further Modules are also available upon request. These are:

- 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 OCRed.

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.



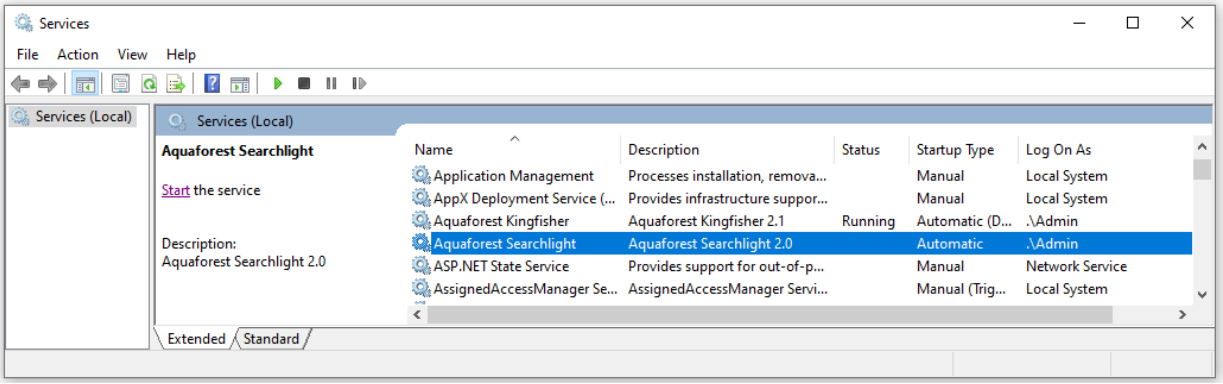
Email support@aquaforest.com to request a key if you do not have one. If you have a valid license key and wish to update it with a new one, go to **Settings > License** tab.

2.3.2 Service Configuration

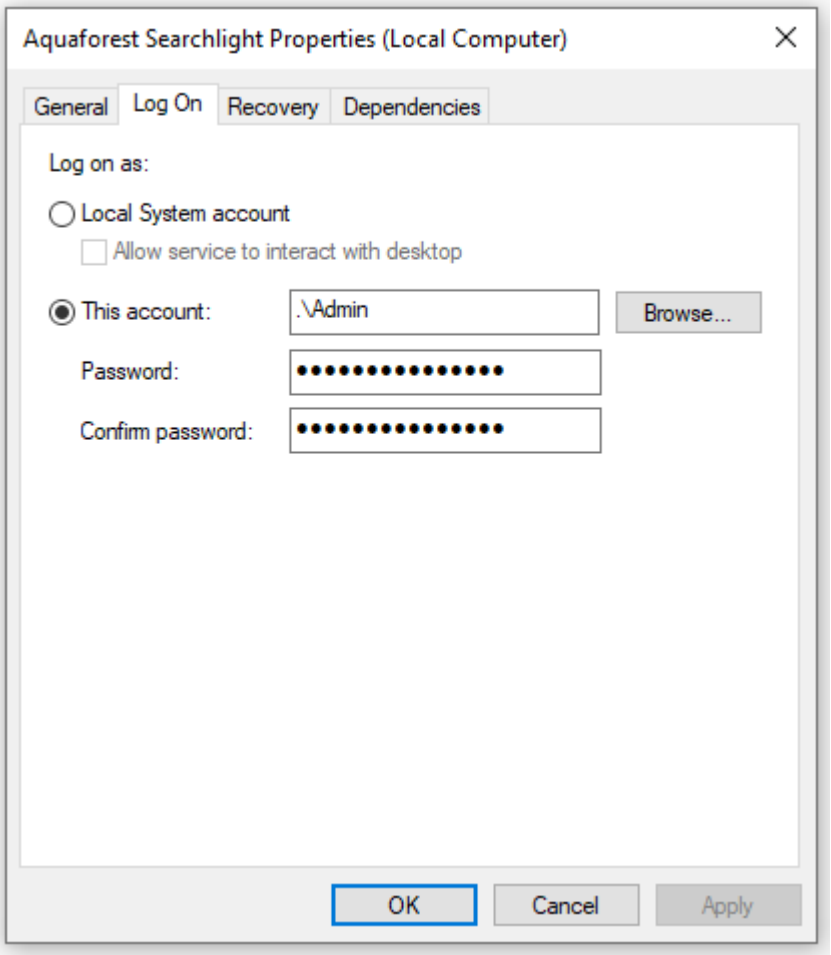
The Aquaforest Searchlight Windows Service is required to log in with an account that has full administrative rights to the File System locations used for Aquaforest Searchlight File System libraries and File System locations used for Errors, Archives and Reports.

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.

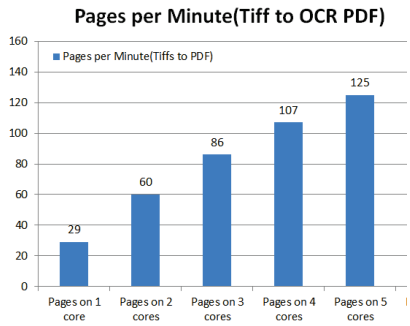
3 Aquaforest Searchlight Modules

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 (Extra Cost)

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

3.2.4 Extended Arabic & Farsi Languages Module (Extra Cost)

Adds support for Arabic and Farsi languages.

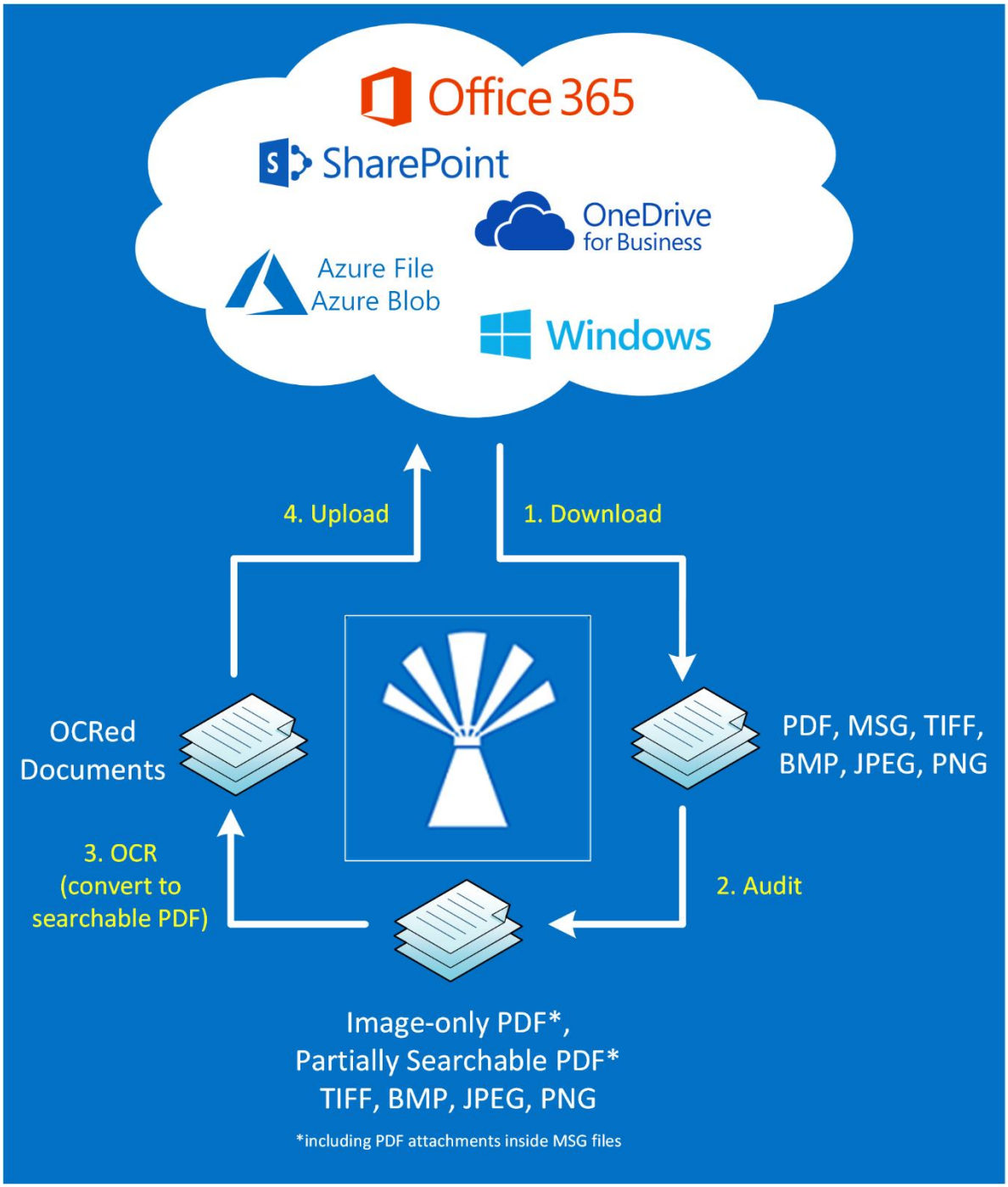
3.2.5 Extended Hebrew Language Support (Extra cost)

Adds support for Hebrew language.

3.2.6 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. 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

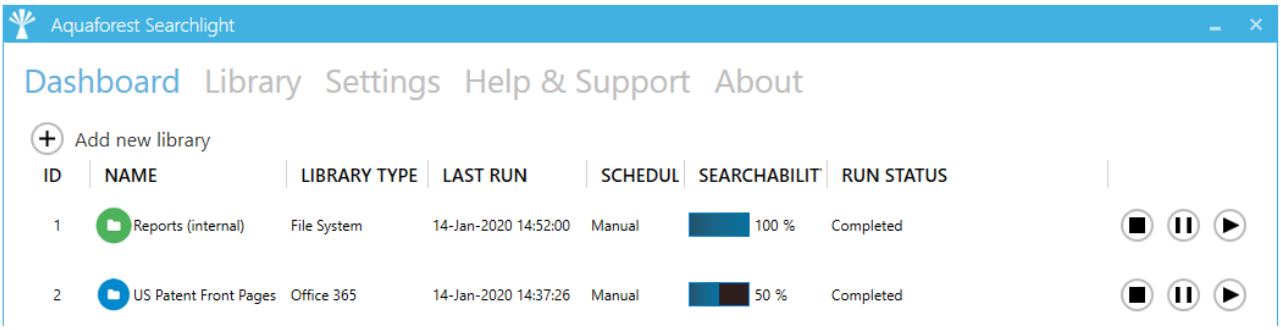
Aquaforest Searchlight currently supports TIFF, BMP, JPG, PNG and PDF documents (including PDF attachments inside MSG files) as input. As a result, candidate documents will always be one of these formats.

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.



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 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/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/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/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 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.

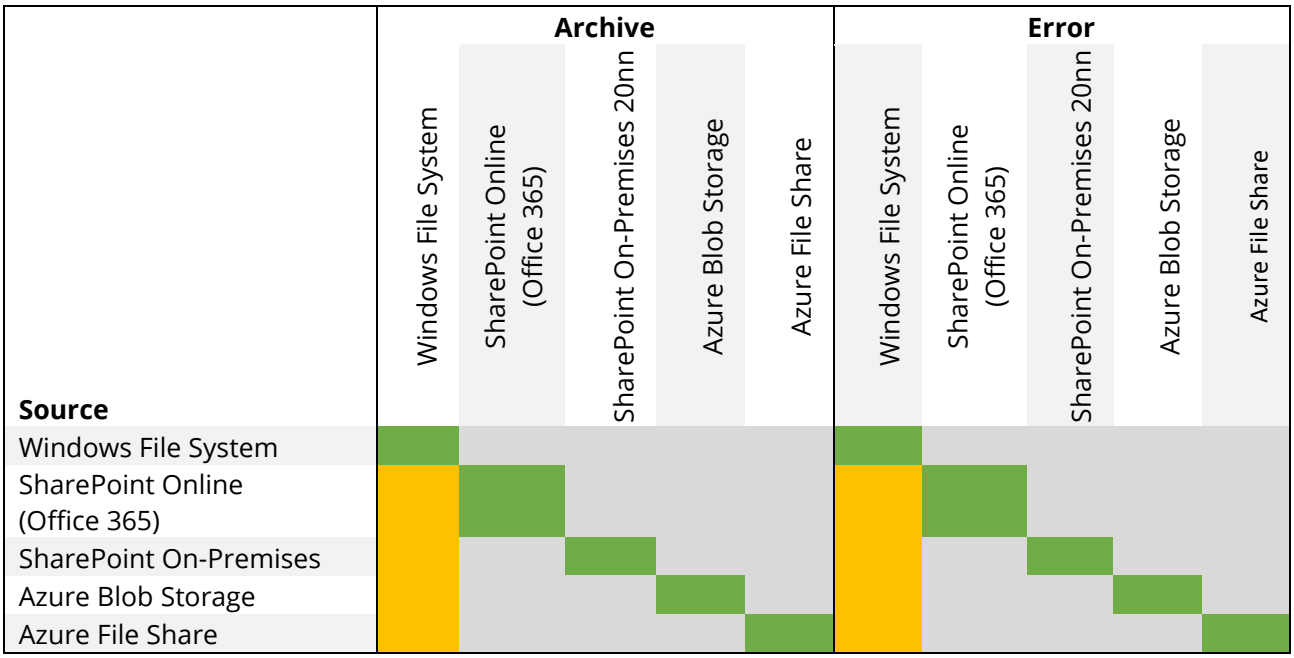
The path is composed of components separated by backslashes (for example \A\B\C\D, each letter is a component), each component can be up to 255 characters in length.

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.



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.

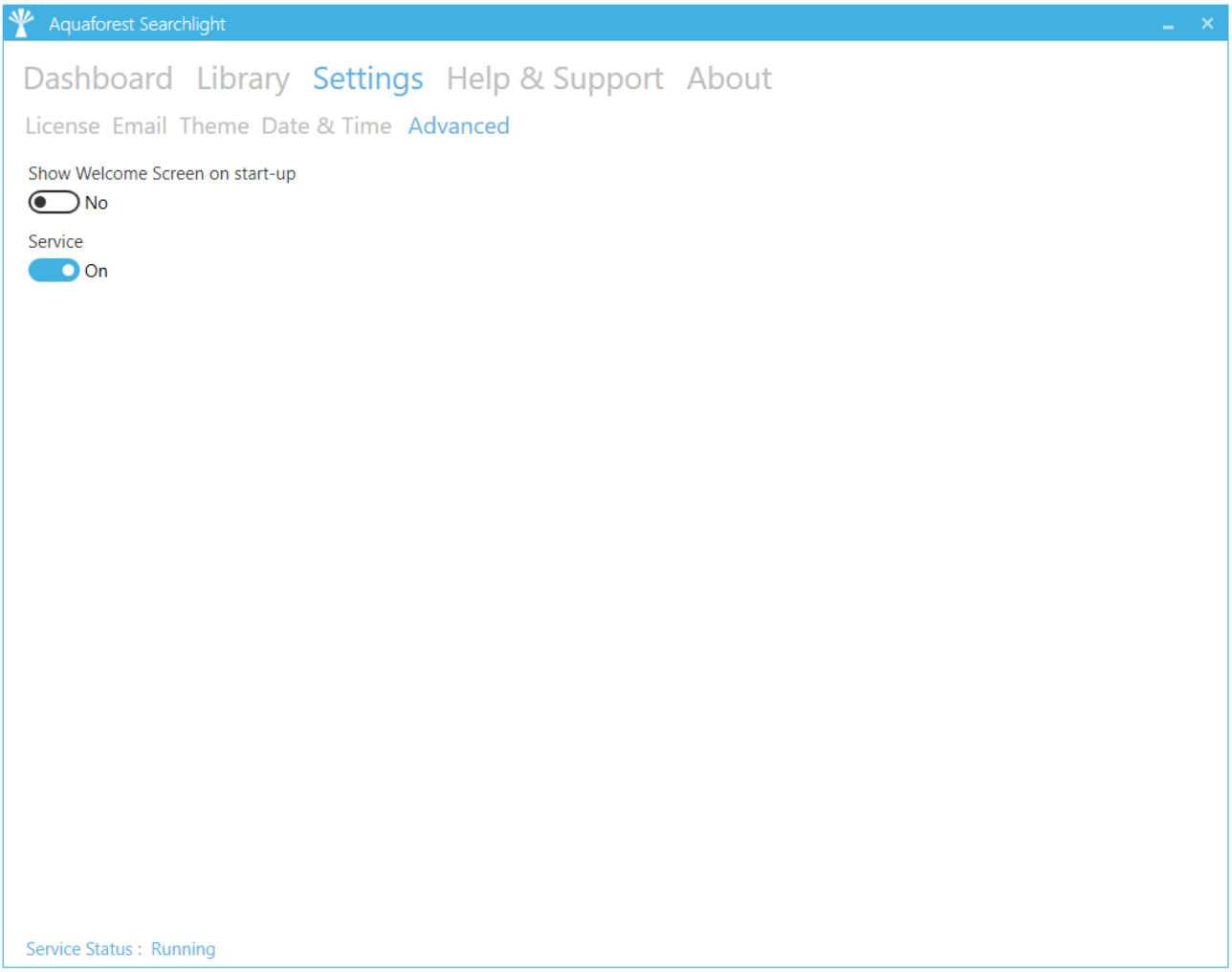
4.6 Archiving

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.

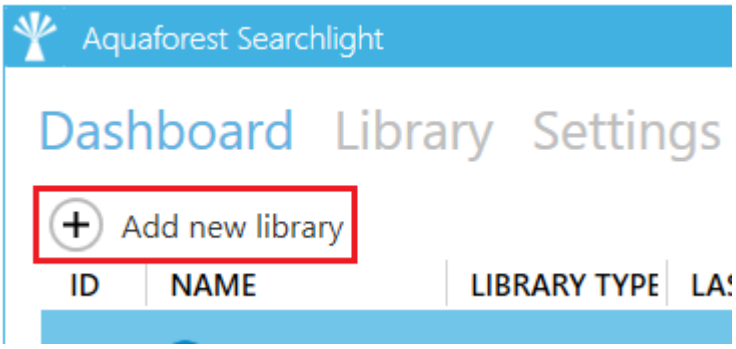
The service can be turned on or off by going to **Settings > Advanced** tab.
The Service Status is displayed at the bottom left of all tabs.



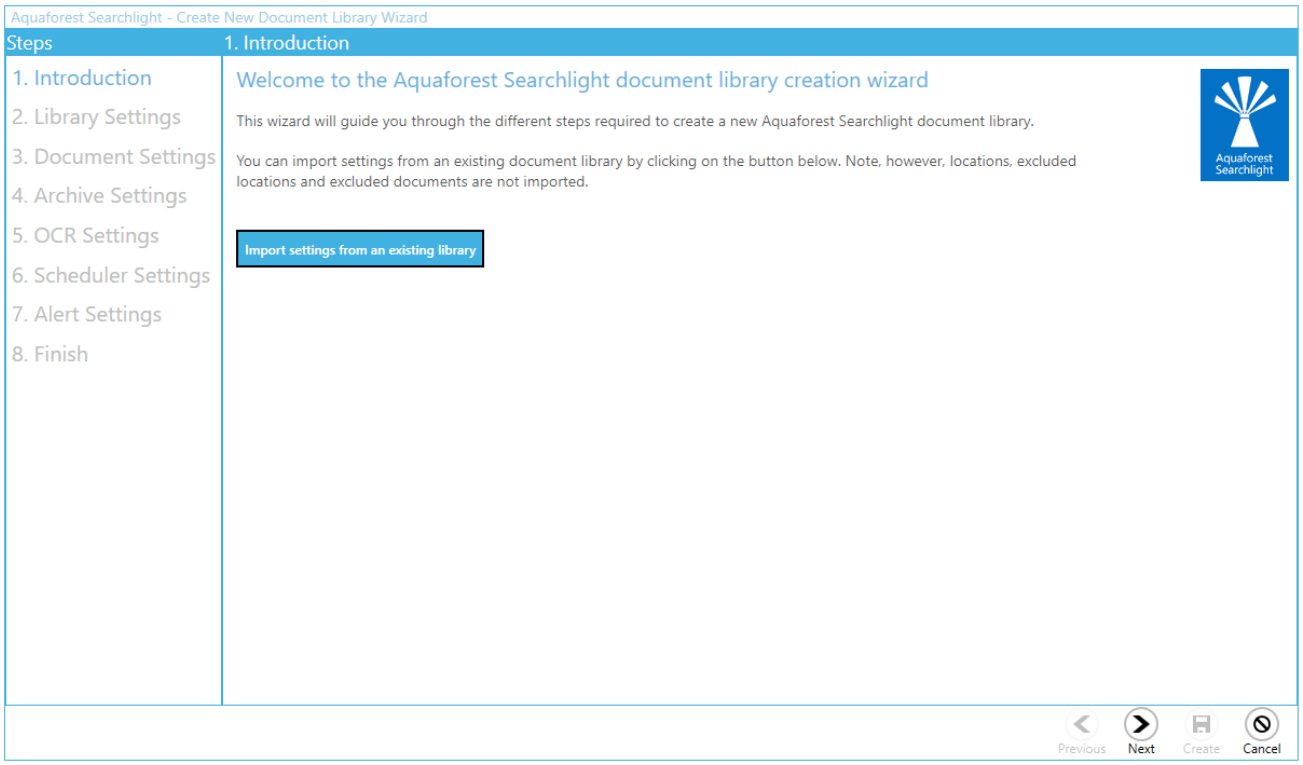
5 Quick Start Guide

5.1 Creating a Library

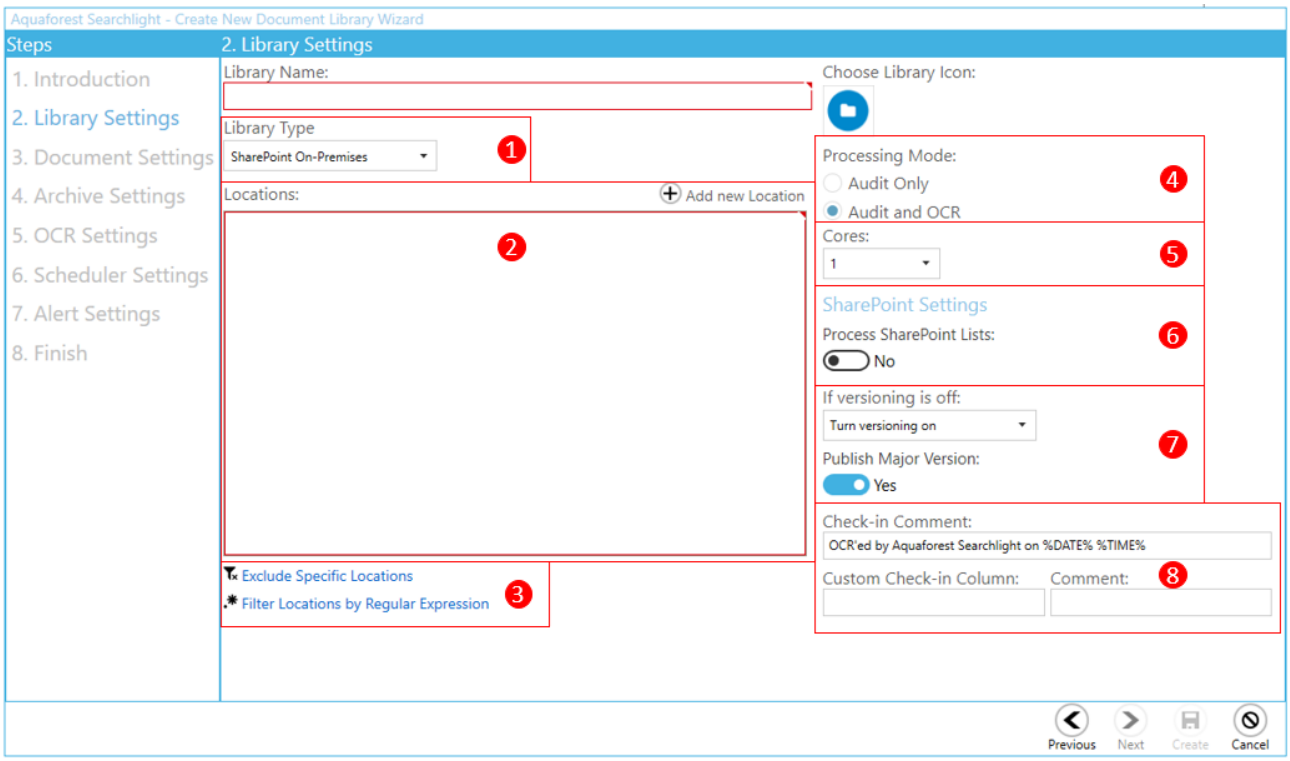
Creating Document Libraries in Aquaforest Searchlight is managed by a wizard. This wizard can be launched by clicking the **Add new library** button on the Dashboard.



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.

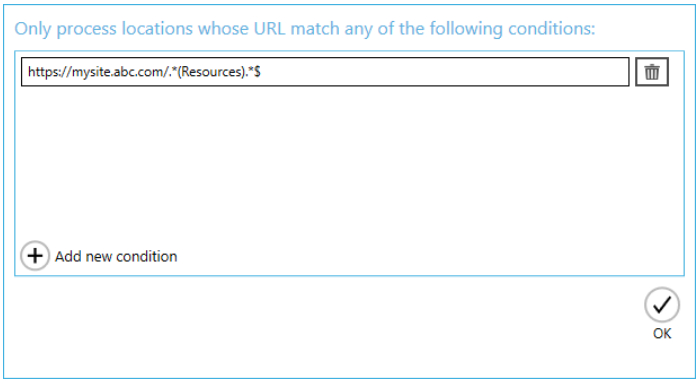


5.1.1 Library Settings



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 URL(s) are excluded.
    b) By regular expressions – locations (URLs) that match the specified regular expressions are included.

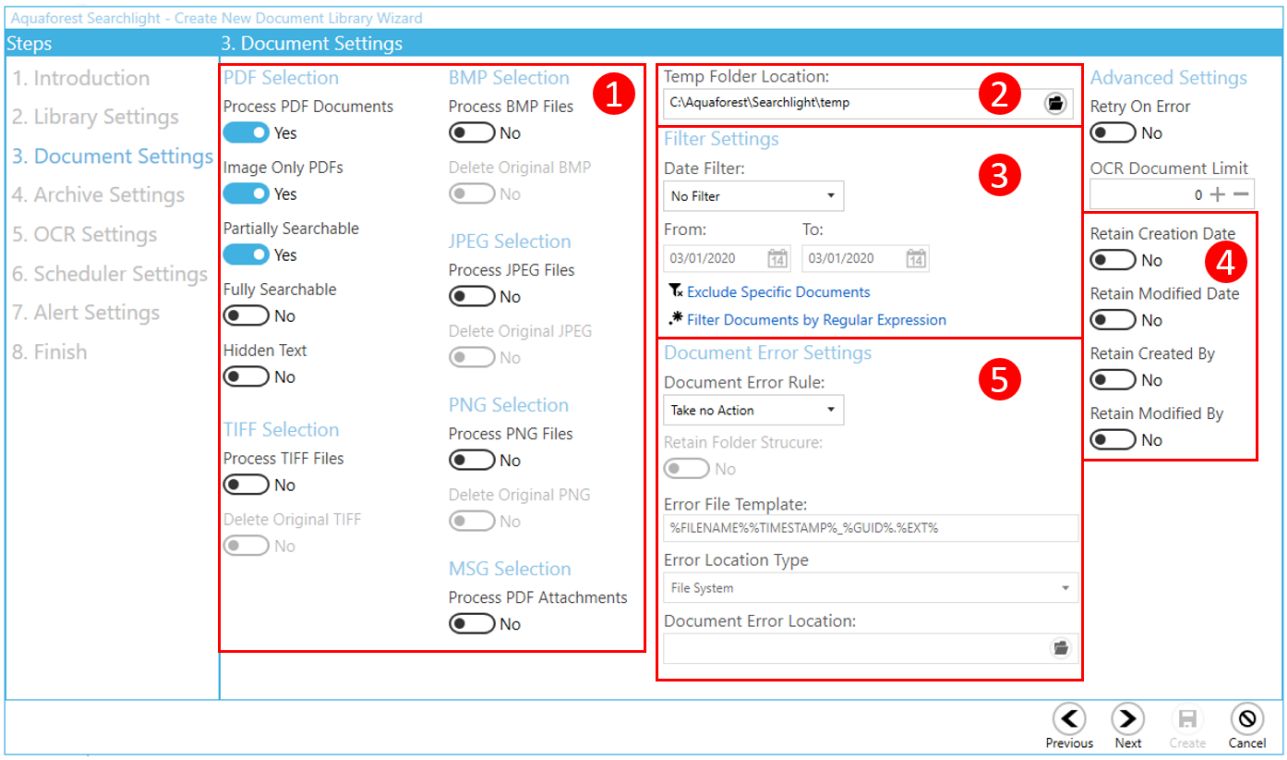
    This is useful if you are processing a whole site collection and want to excluded specific locations and/or include only specific sites or libraries. For instance, you may want to only process sites and libraries containing the word “Resources” in their URL:



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 OCRed 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

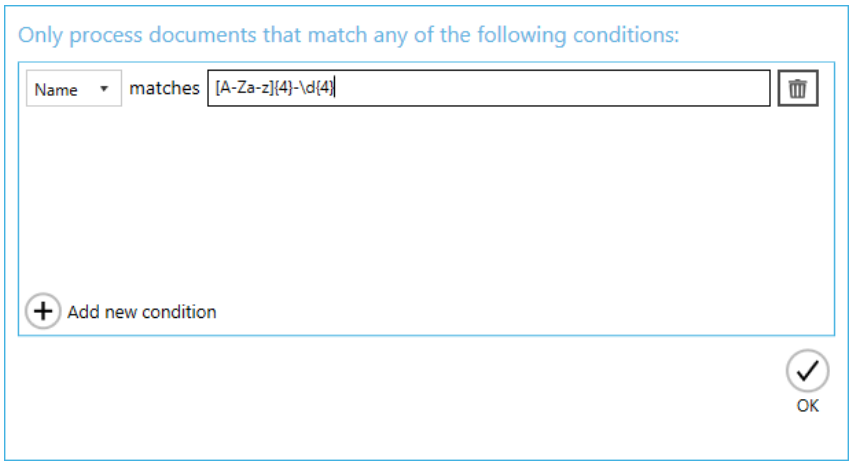
This page enables the user to specify rules and criteria for the selection of documents to be processed.



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

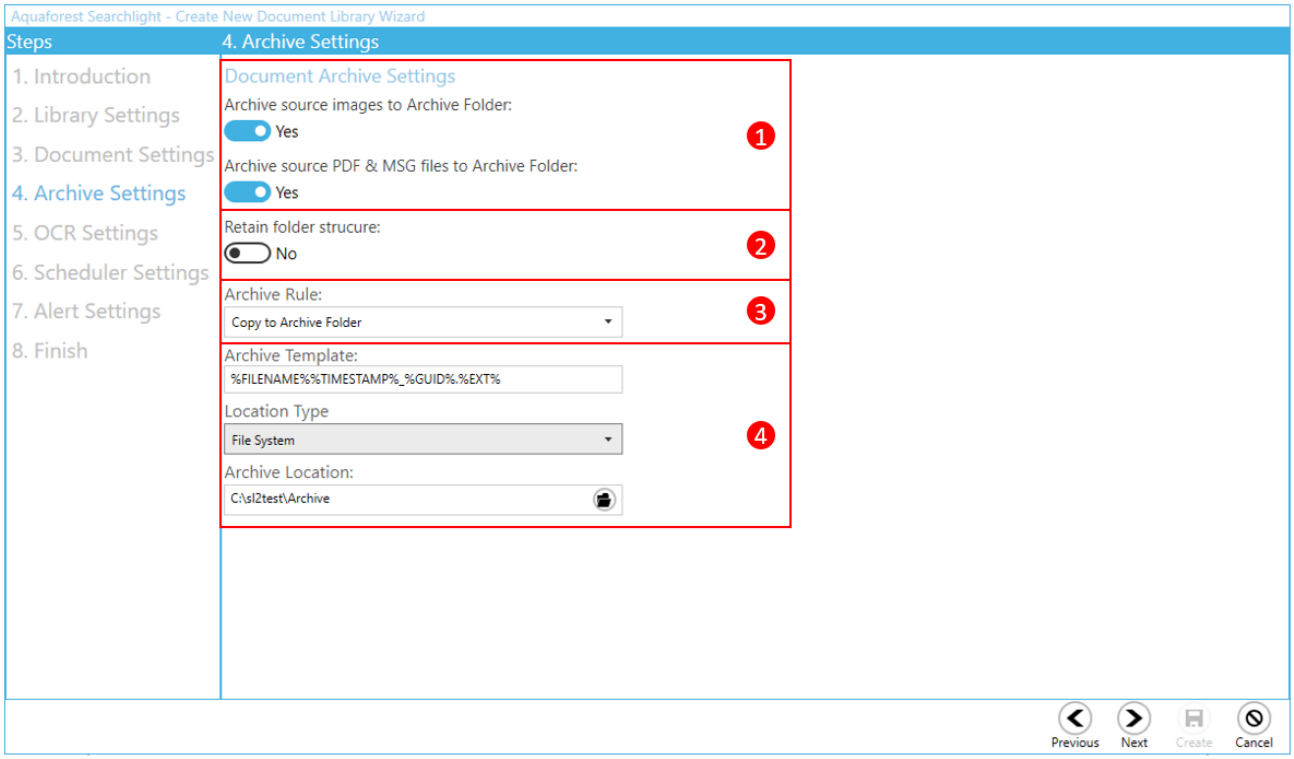
    For instance, you may want to only process documents with the name format "ABCD-1234":



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.

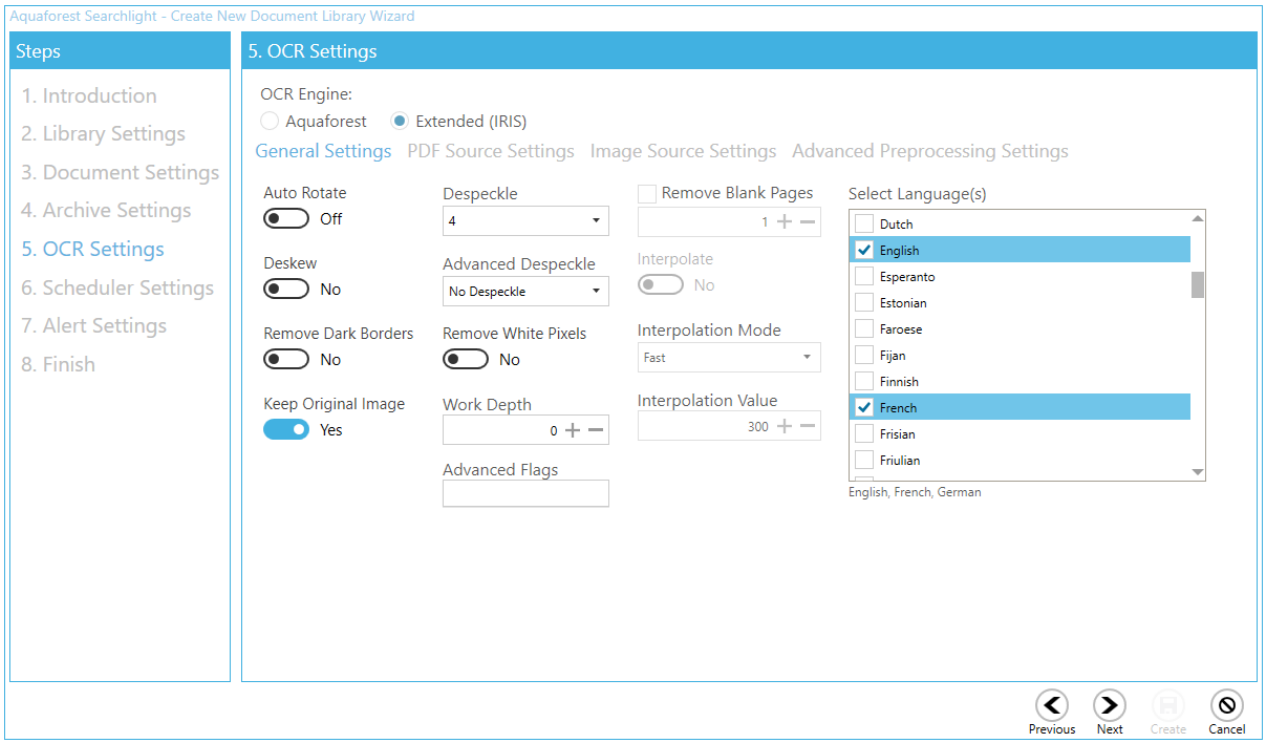


1. Select whether you want to archive just image files (TIFF, BMP, JPG and PNG) or PDF and MSG files.
2. Select if you want to retain the existing folder structure within the archive.
3. Select the archive rule – Copy to Archive Folder
4. Select the archive filename format, storage type (File System,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

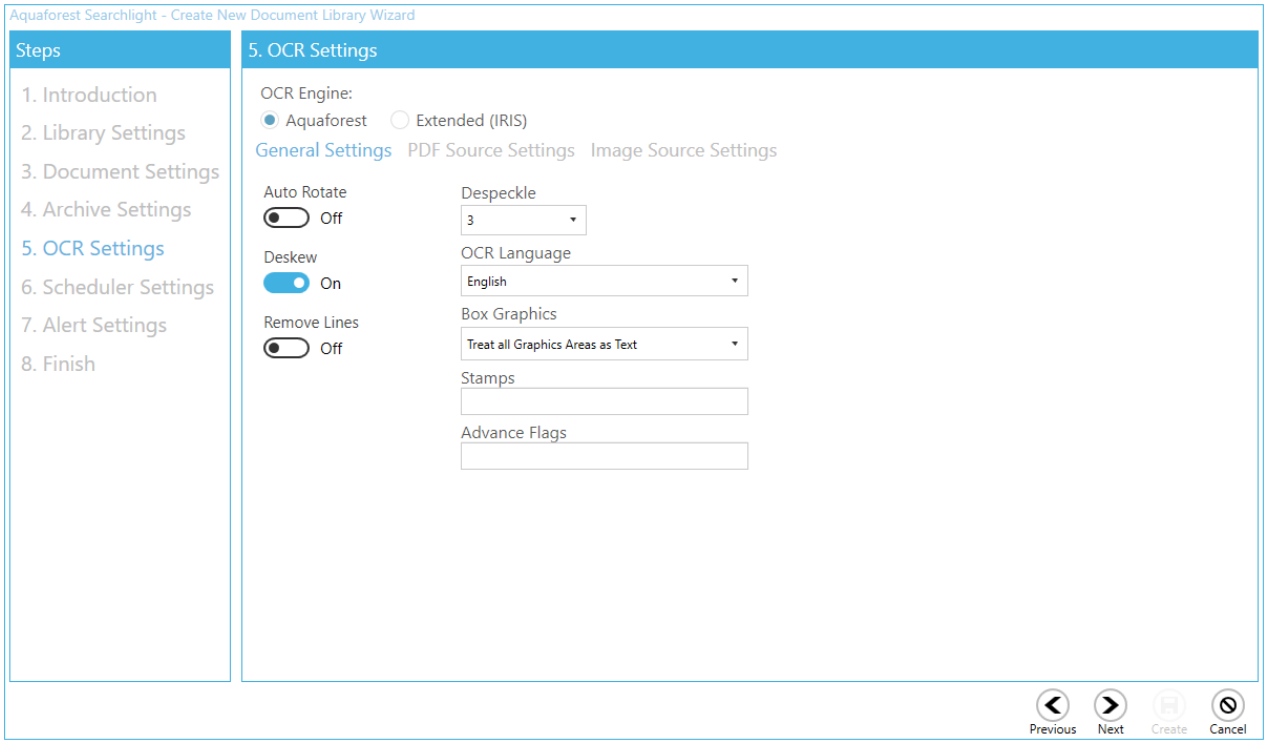
5.1.4 OCR Settings

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

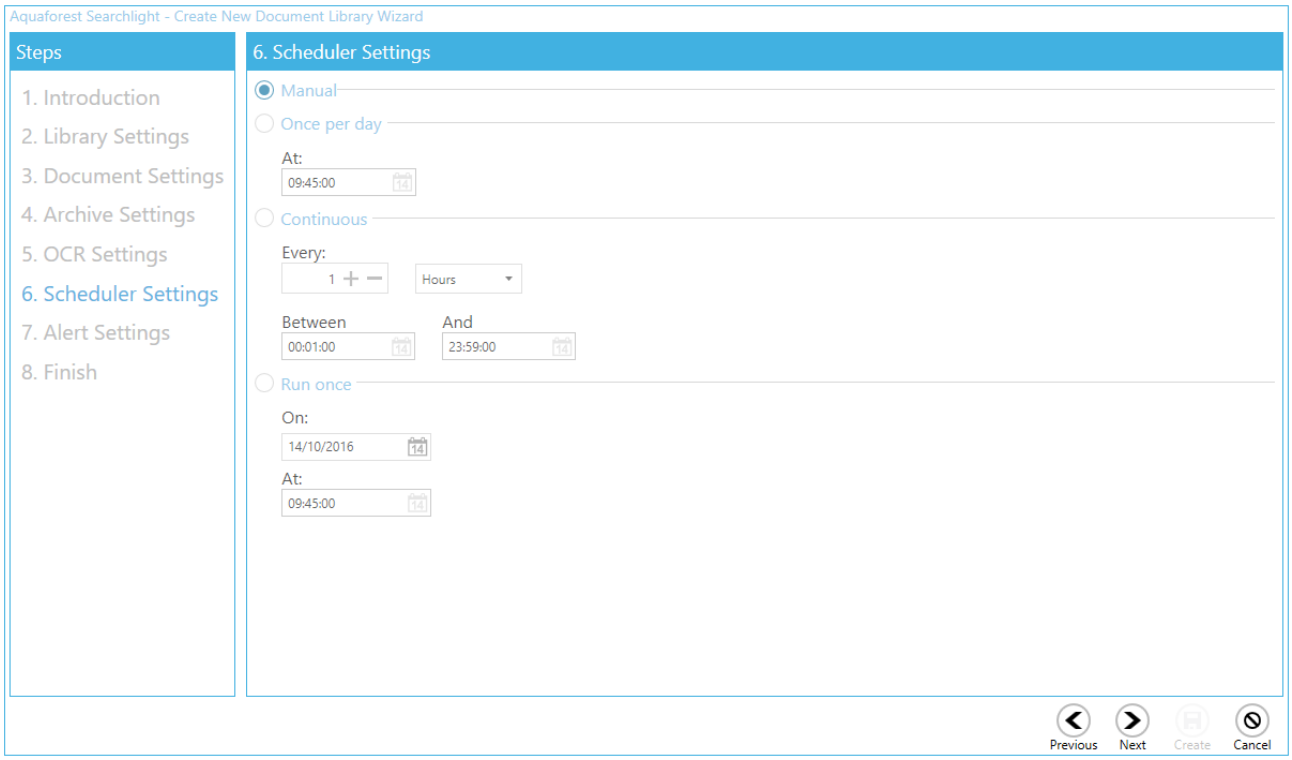


5.1.4.2 Standard OCR Engine Settings



5.1.5 Scheduler

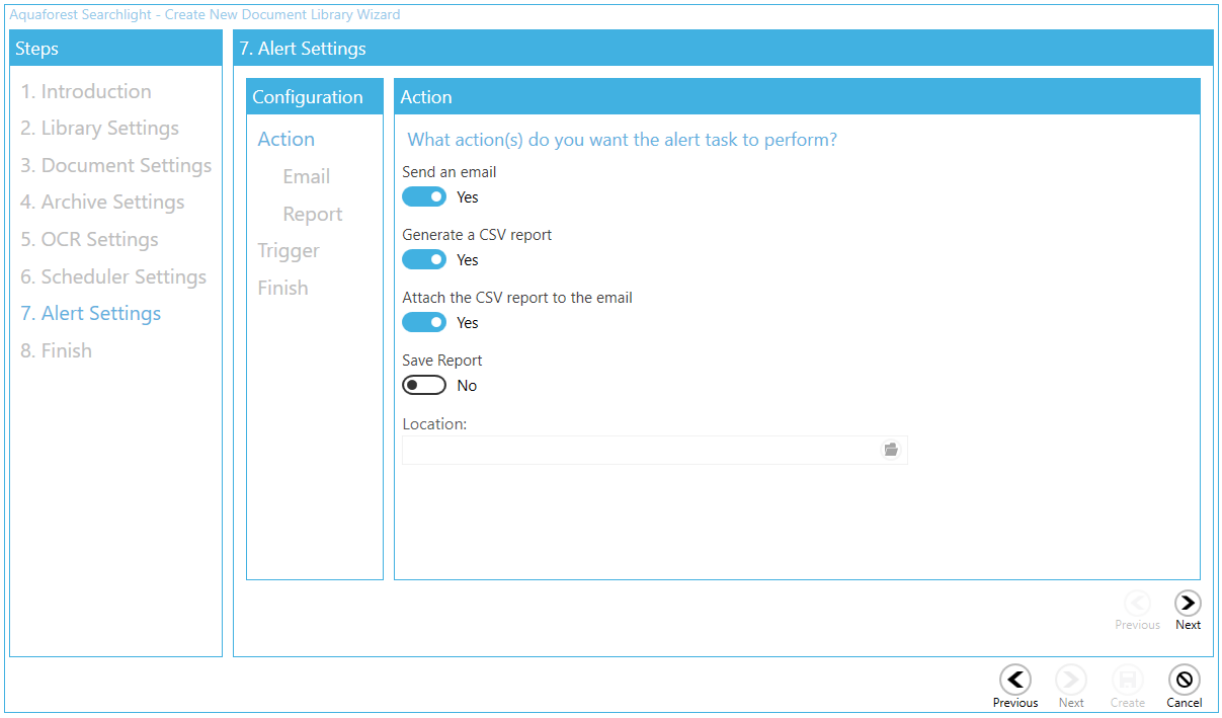
The scheduler allows Aquaforest Searchlight to automate the running of document libraries. You can either run it manually, or run periodically, every day at a specified time or every hour etc.



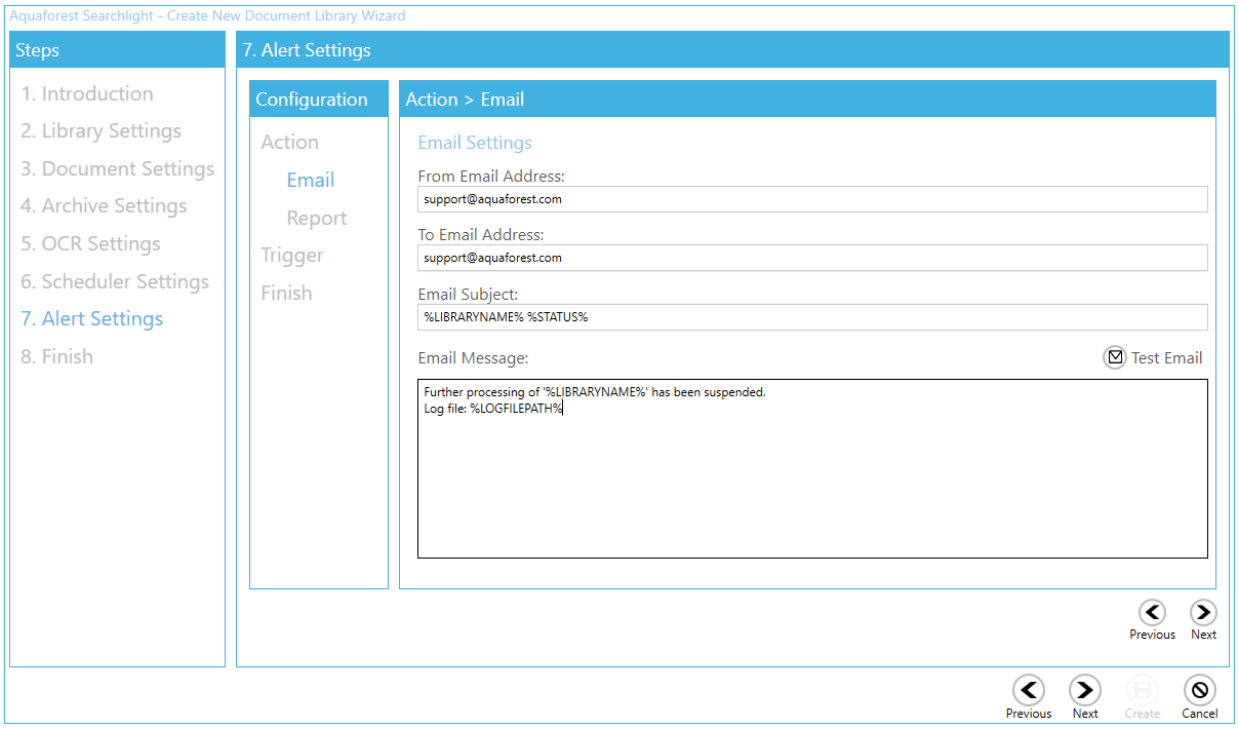
5.1.6 Alert Settings

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.

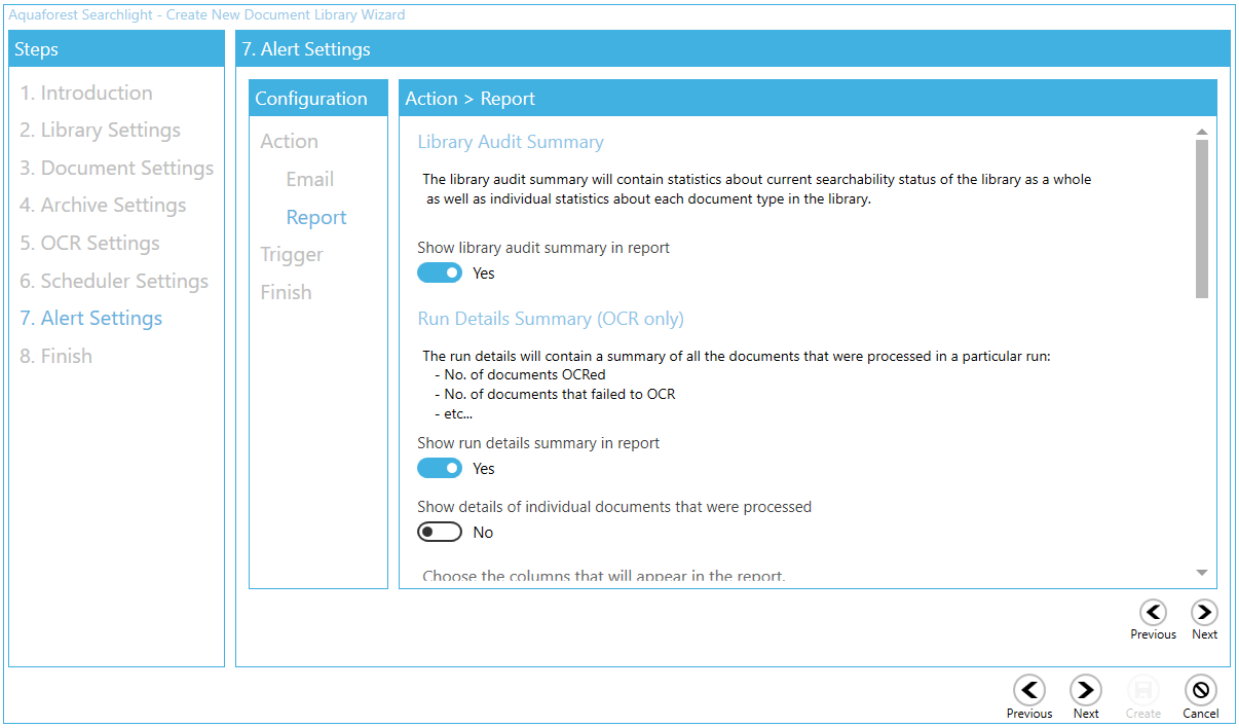
1. Select the action(s) you want to perform.



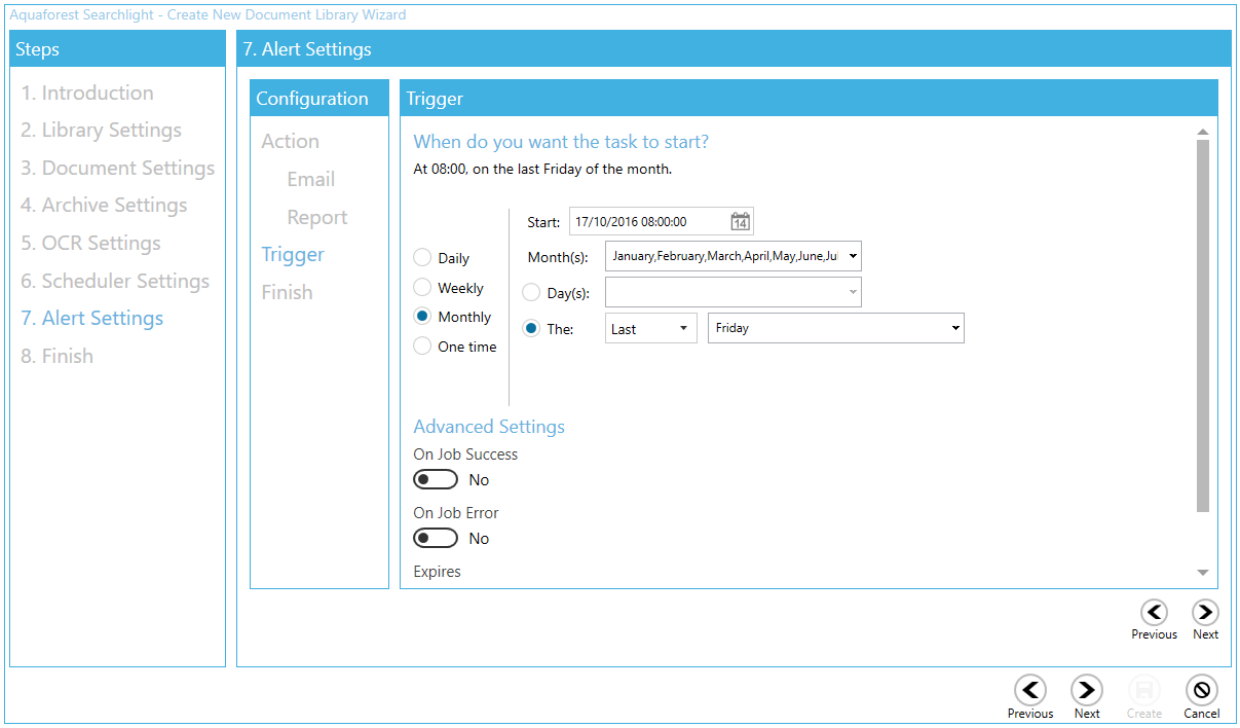
2. Select the email settings.



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.

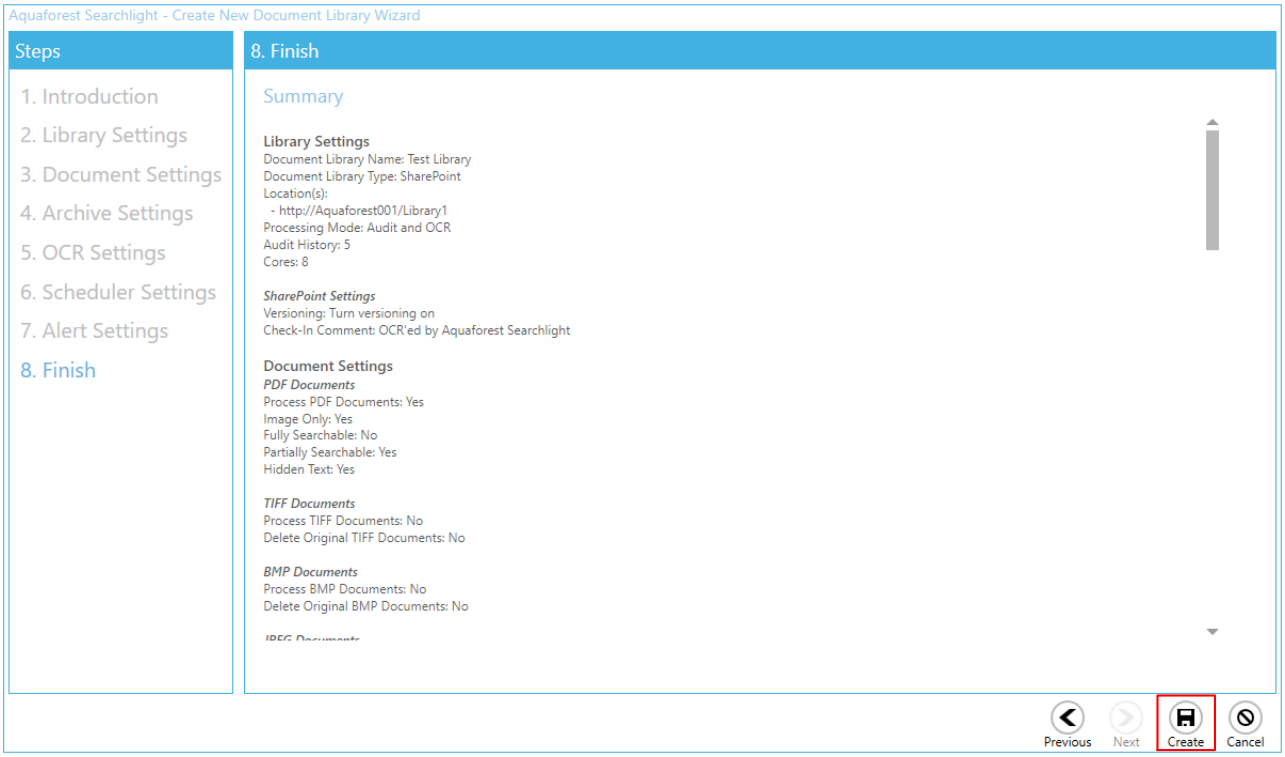


4. Select when you want the task to run. Based on the current settings, you will get an email with the report attached sent to the recipient every last Friday of the month at 8 am.

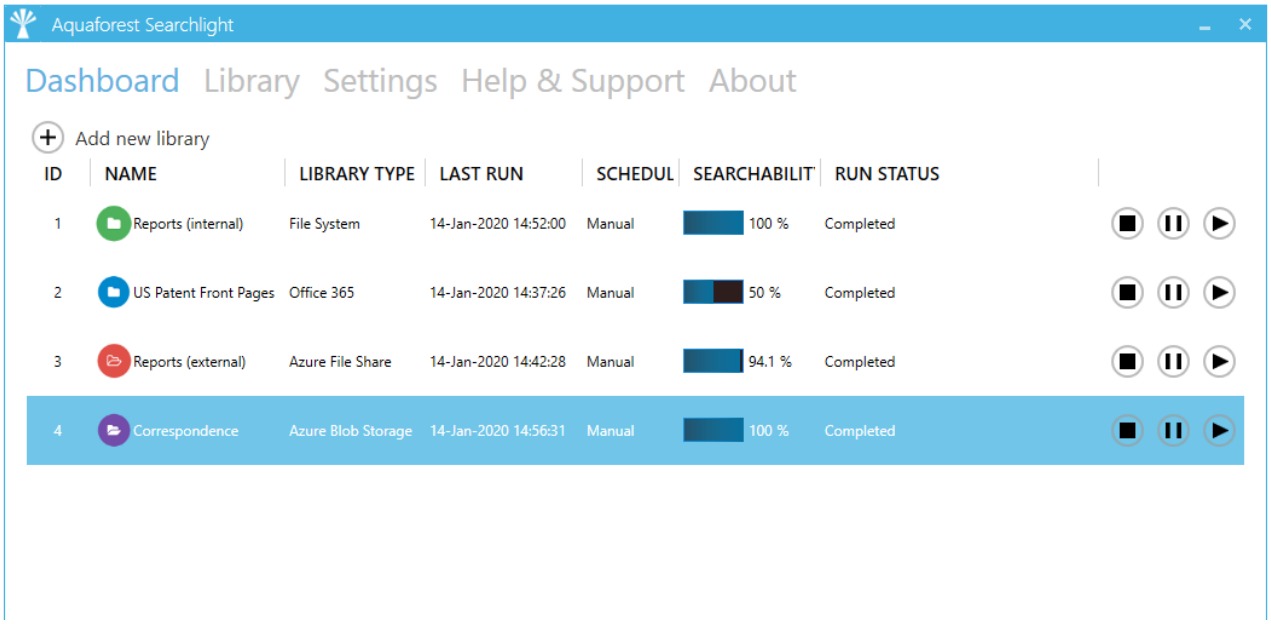


5.1.7 Finish

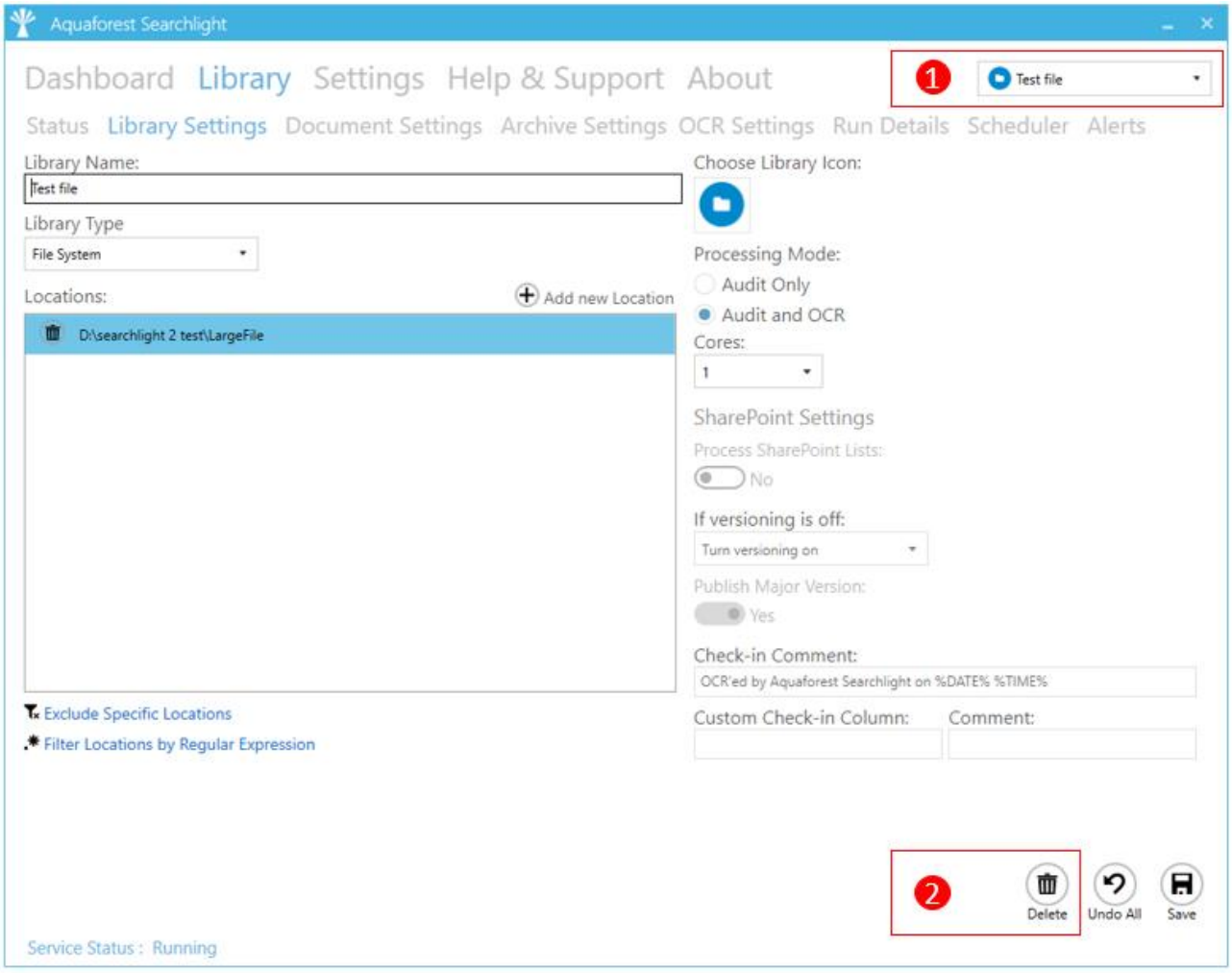
On the **Finish** page, you will get a summary of all the settings you selected for this library. You can review them to see if you missed anything. If not, click on the **Create** button at the bottom of the wizard to create the library.



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

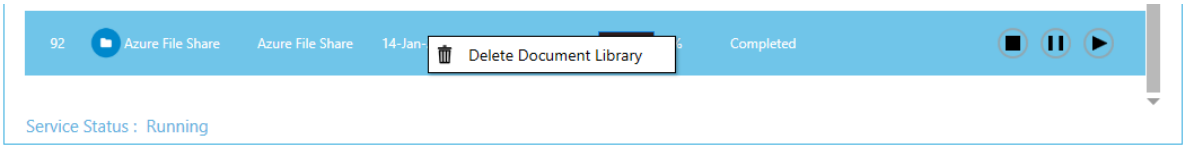


All the settings of a library can be edited by double-clicking the library from the dashboard, or by selecting the library and clicking on the Library Tab.

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**

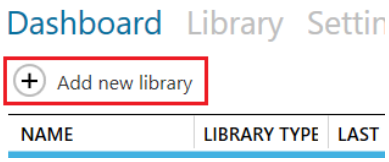


5.3 Importing settings from an existing Library

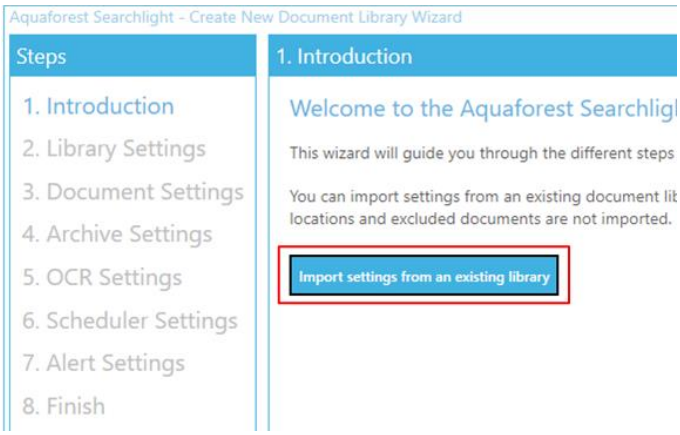
Searchlight also provides the ability to import settings from an existing library. However, "locations", "excluded locations" and "excluded documents" are not imported because it is not allowed to have the same locations in multiple libraries.

To import settings:

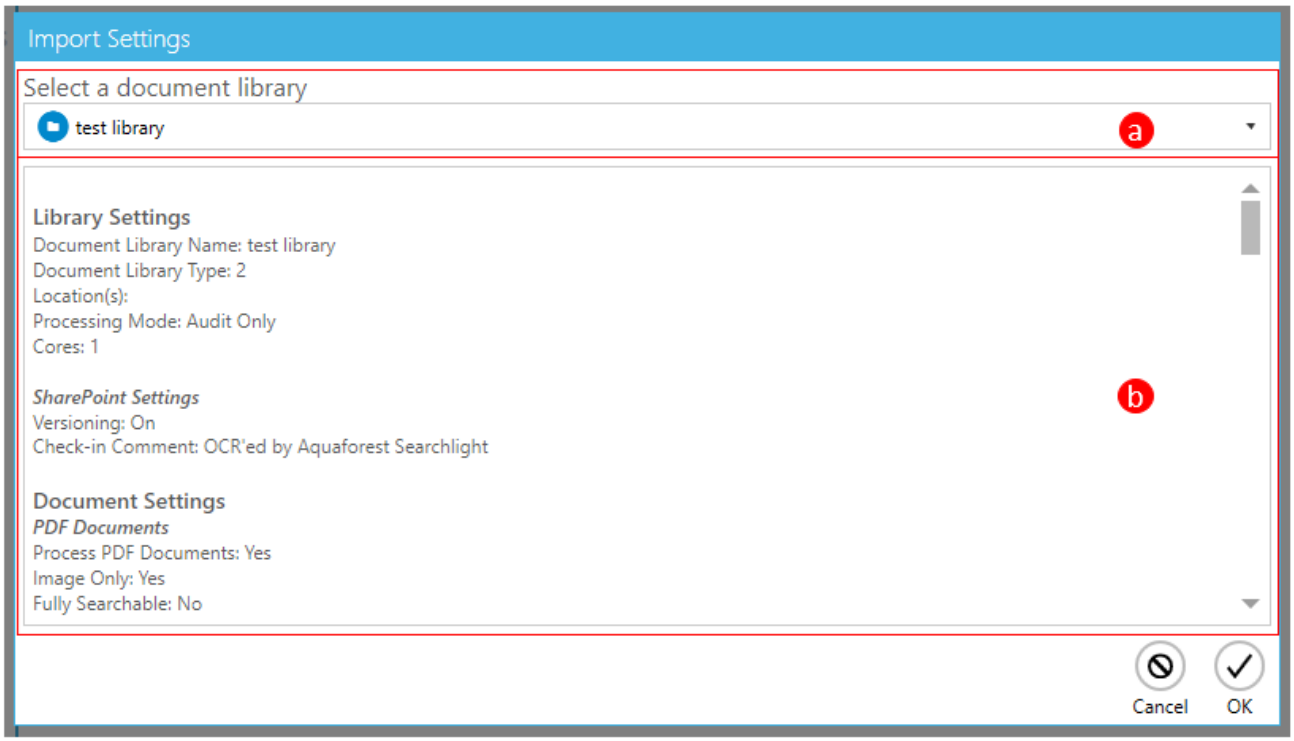
1. Click on the **Add new library** button on the Dashboard to open the wizard.



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



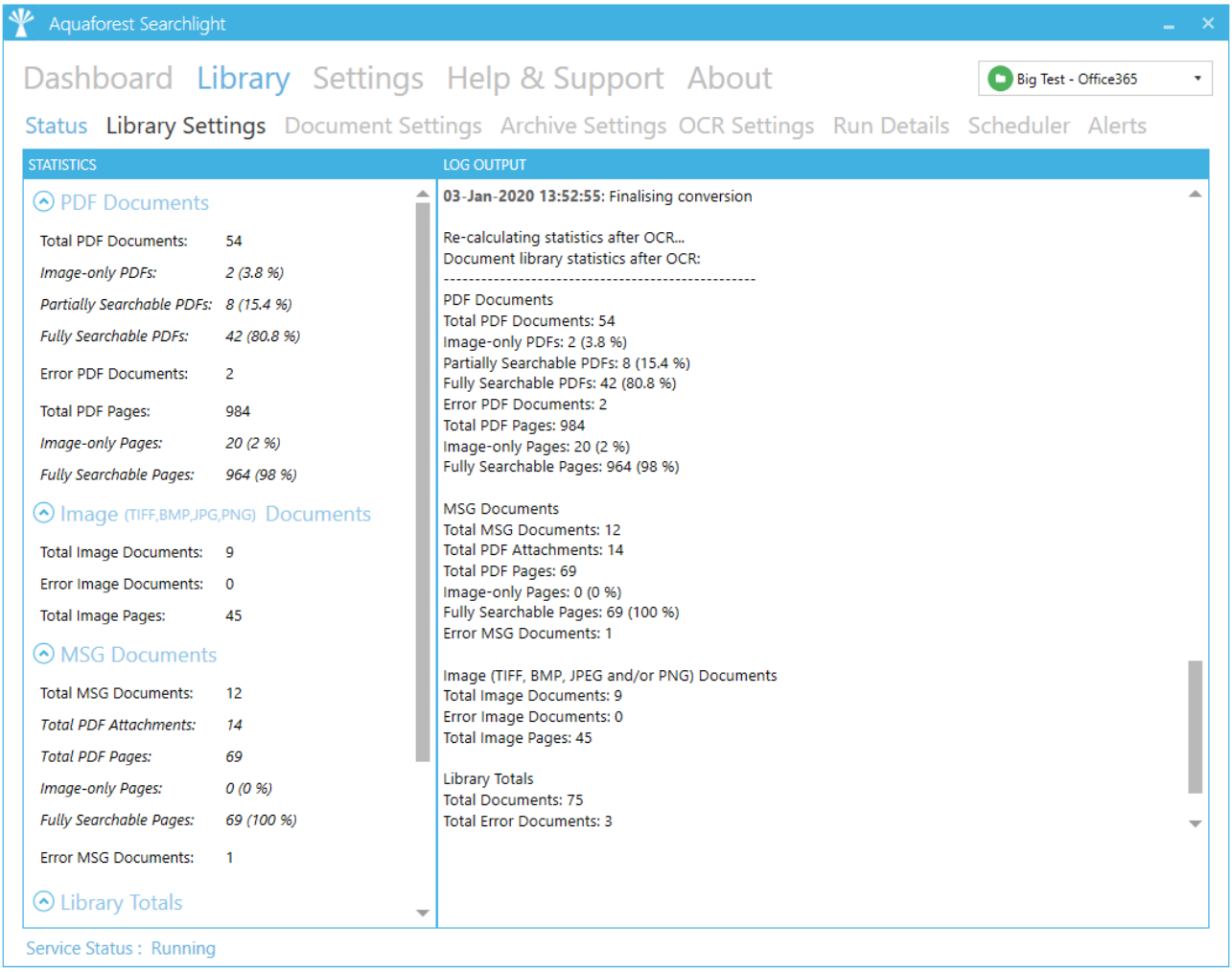
3. From the Import Settings window:
   a. Select the document library from which the settings are to be imported.
   b. A summary of the settings of the selected document library will be displayed in the text box underneath.



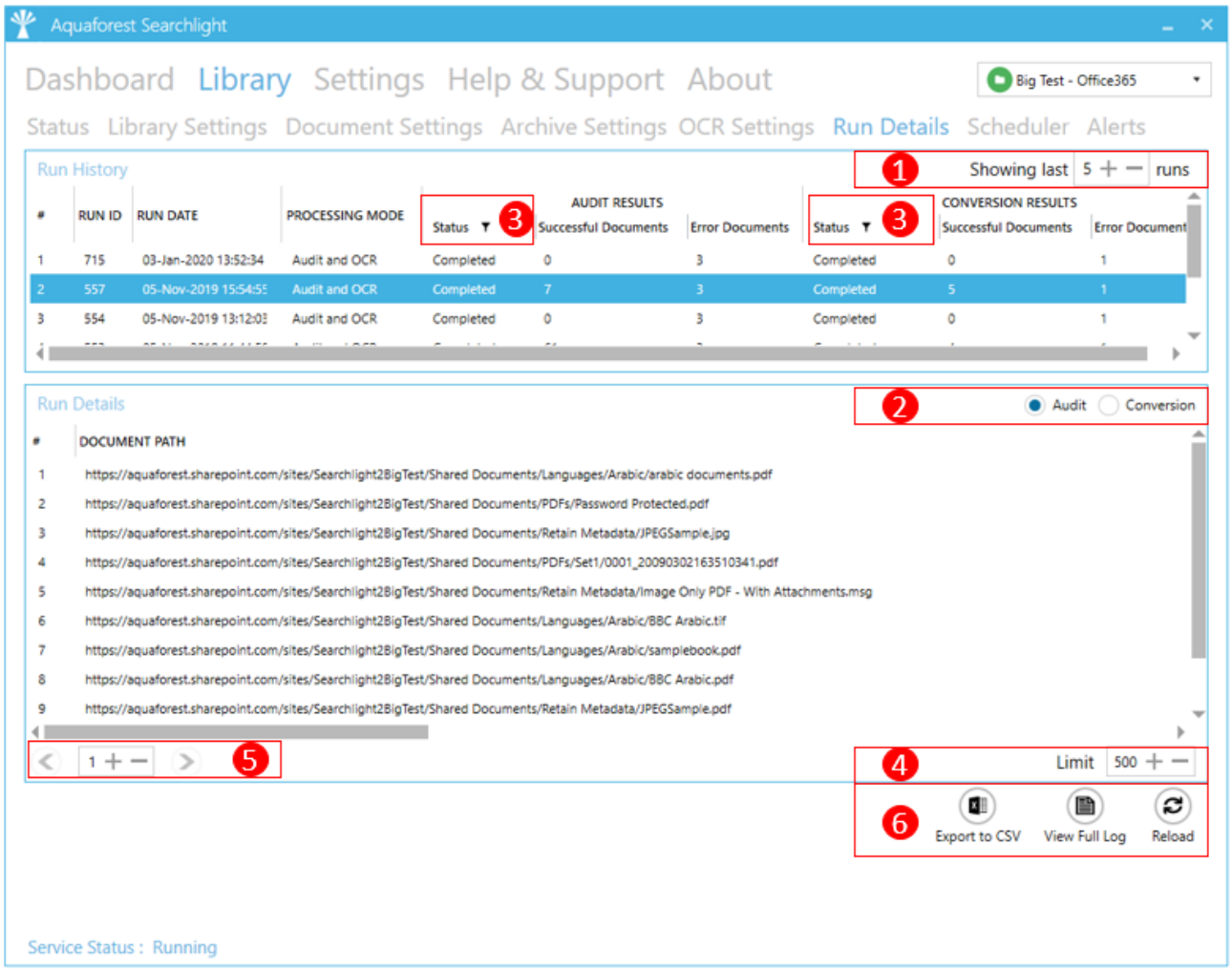
4. After clicking **OK** from the Import Settings window, go to **Library Settings** and add the location(s) to process. Optionally, add specific locations and documents to exclude.
5. Review all the settings in the other sections and click **Create**.

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:



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.

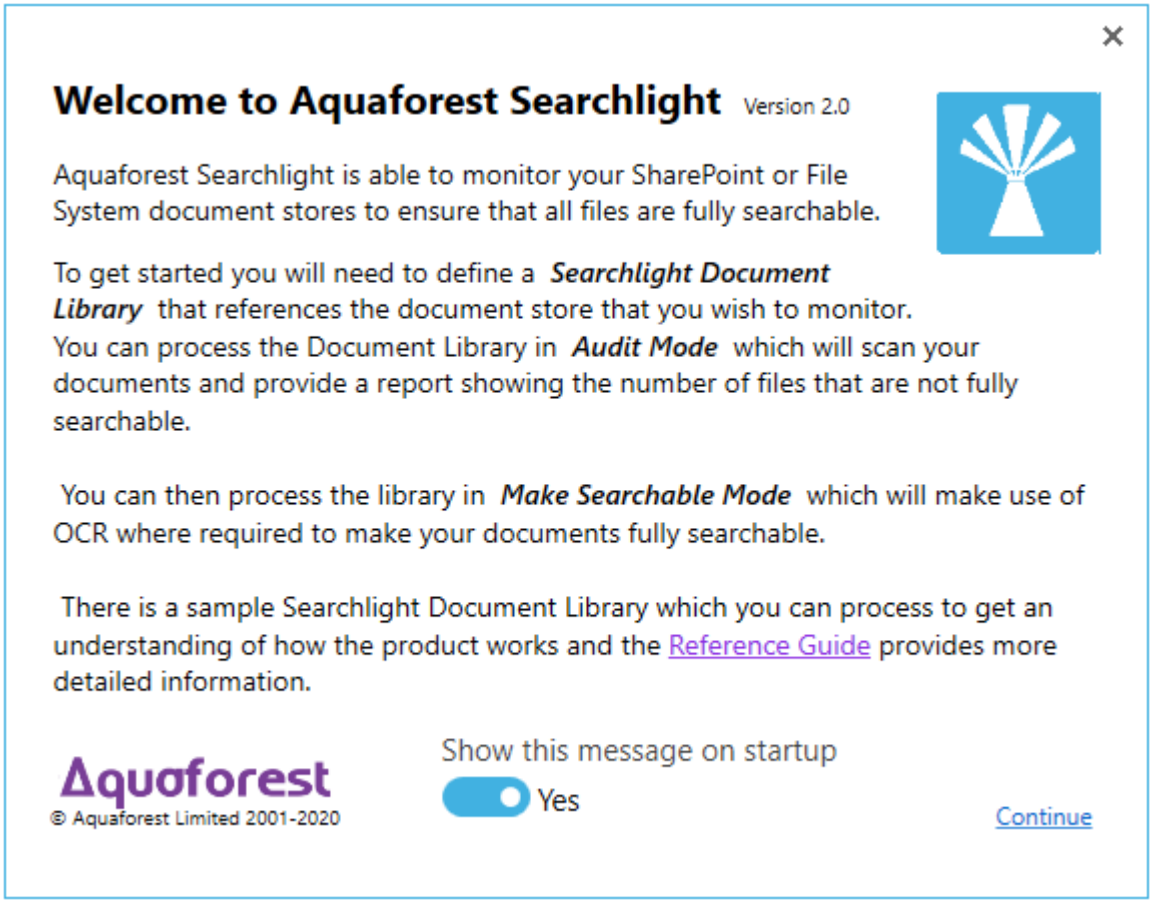


1. Select the number of previous runs to show. You need to click on the **Reload** button after updating this value. Clicking on a run history will display its details in the **Run Details** section below.
2. Select whether you want to display the documents that were audited or OCRed for that specific 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).

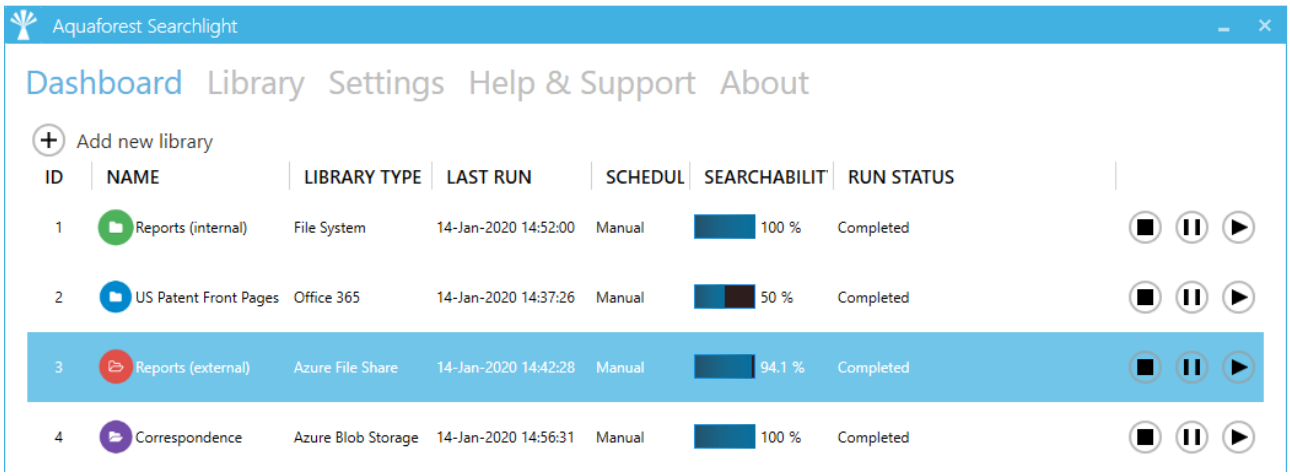
6 The Aquaforest Searchlight Tool

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.



6.2 Dashboard



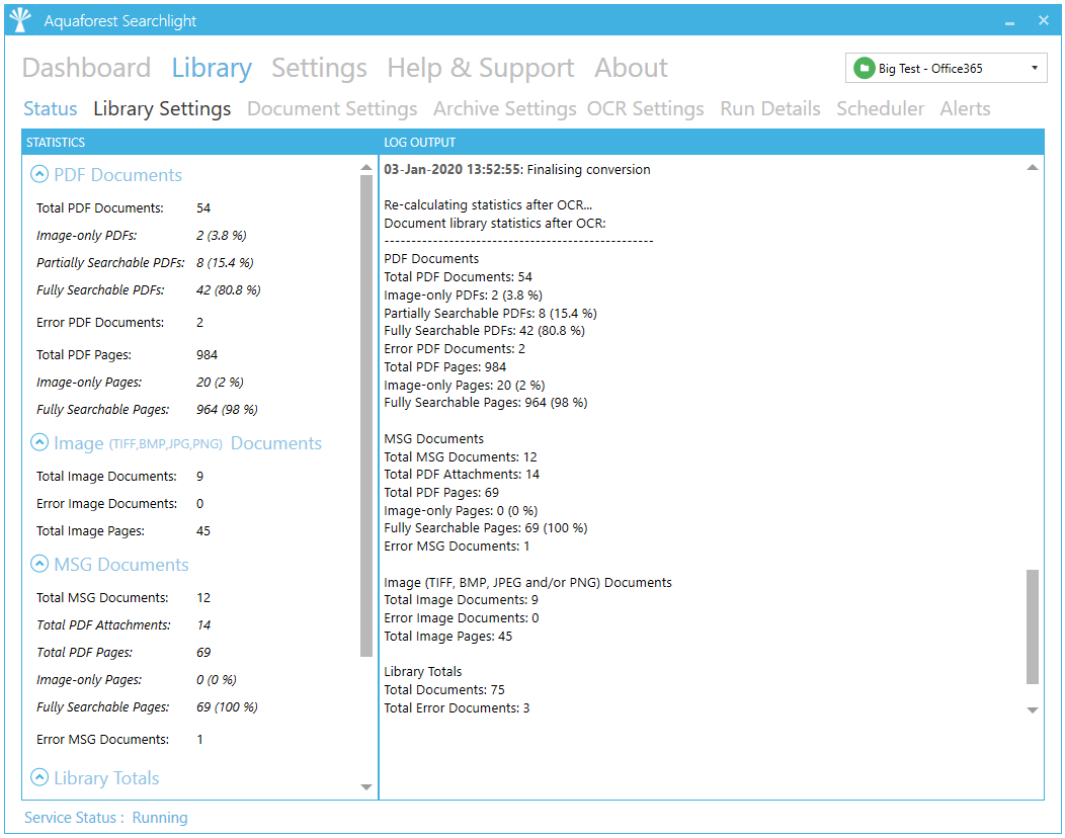
The dashboard gives a summary of the status of all the document libraries that have been created by the user.

Column	Description
Name	Name of the document library
Library Type	The type of the document library: • SharePoint On-Premises • SharePoint Online (Office 365) • File System • Azure Blob Storage • Azure File Storage
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: • Running • Completed • Error • Aborted
■ ❚❚ ▶	Abort, Pause, Start

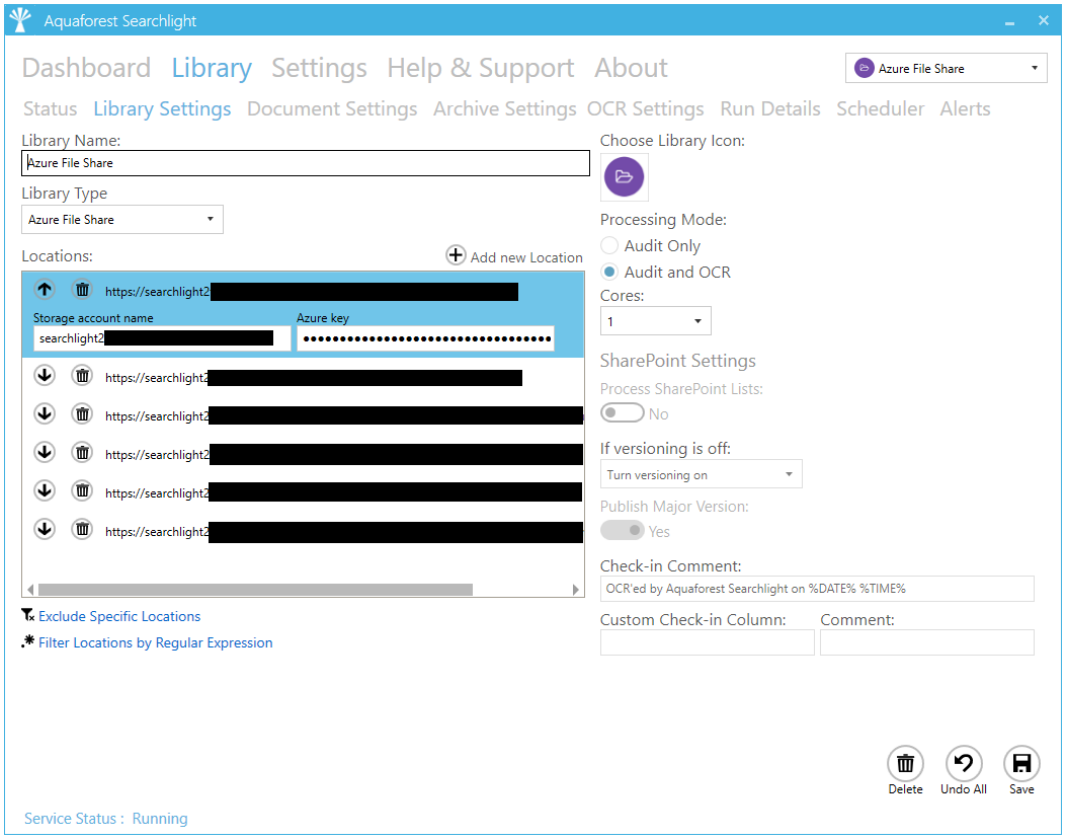
6.3 Library

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.



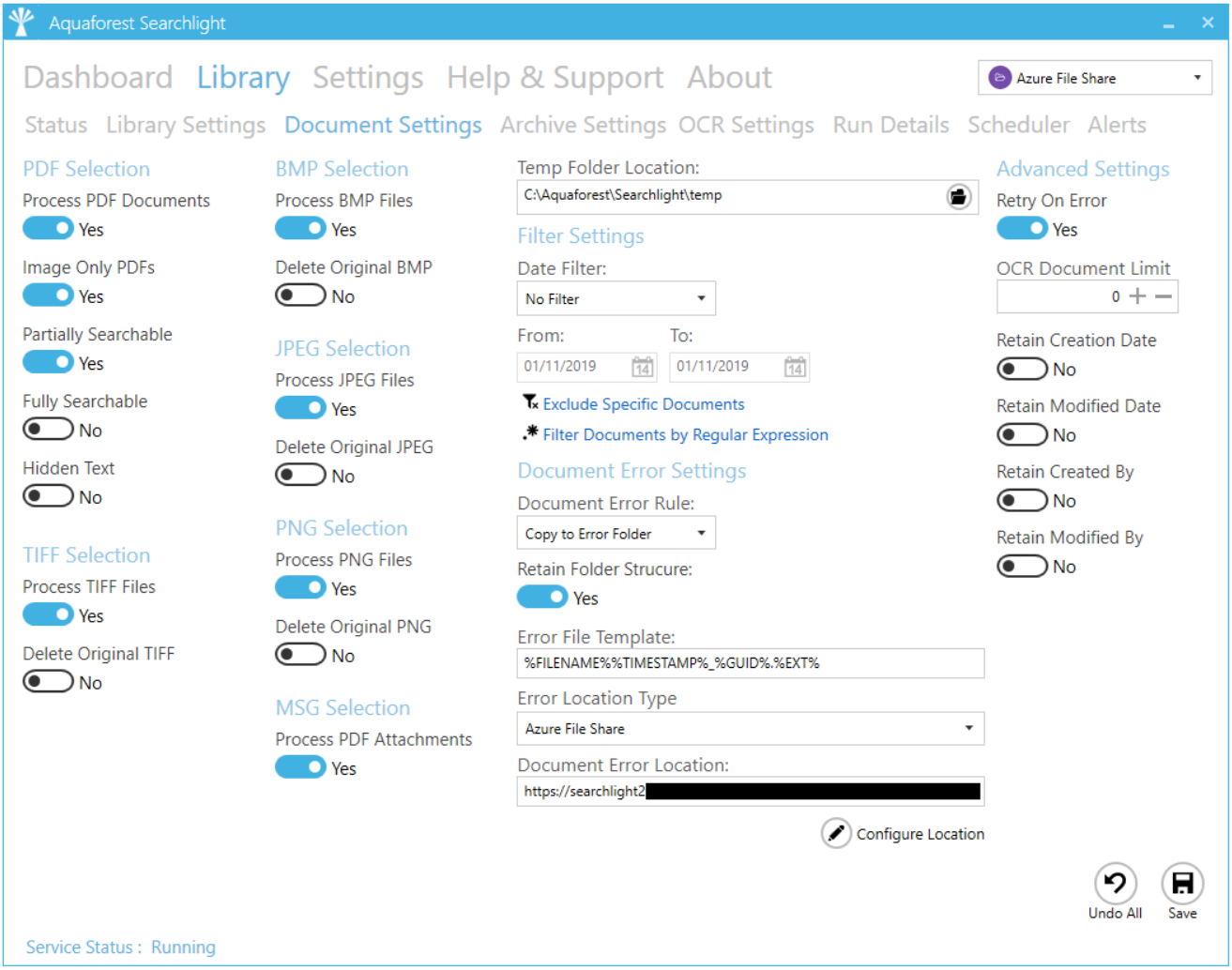
6.3.2 Library Settings



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	• 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.
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	The check-in comment applied to the updated SharePoint file version. There is also the option of specifying the following templates in the check-in comment: • **%DATE%** : will be replaced by the date the document OCRed • **%TIME%** : will be replaced by the time the document OCRed
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	The comment to add to the Custom Check-in Column. There is also the option of specifying the following templates in the comment: • **%DATE%** : will be replaced by the date the document OCRed • **%TIME%** : will be replaced by the time the document OCRed

6.3.3 Document Settings



Setting	Description
Process PDF	Whether or not to process PDF documents
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.

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	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: • Copy to error folder • Move to error folder (for file system library type only)

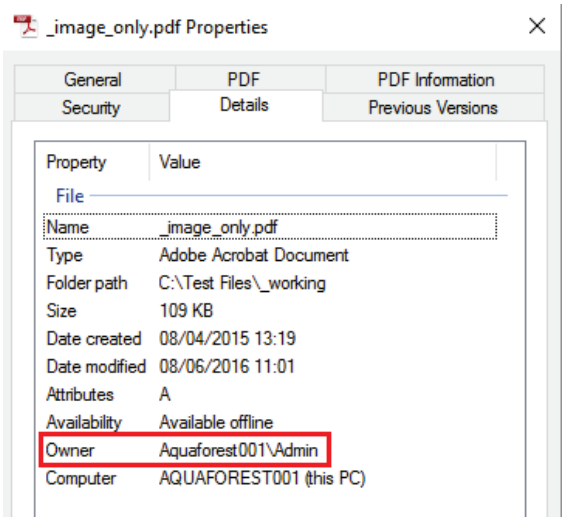
Setting	Description
Retain Folder Structure	Option to retain document's folder structure when copied to error location
Document Error Location	The path of the error location
Document Error Location Type	File System SharePoint Office 365 Azure Blob Storage Azure File Storage
Retry	Whether or not to re-process 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

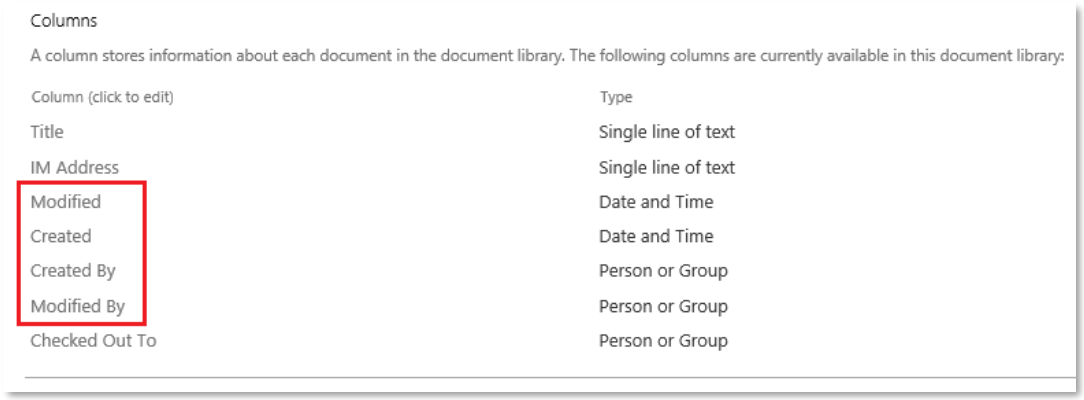
- * "Create User" maps best to "Owner" in Windows File System metadata.



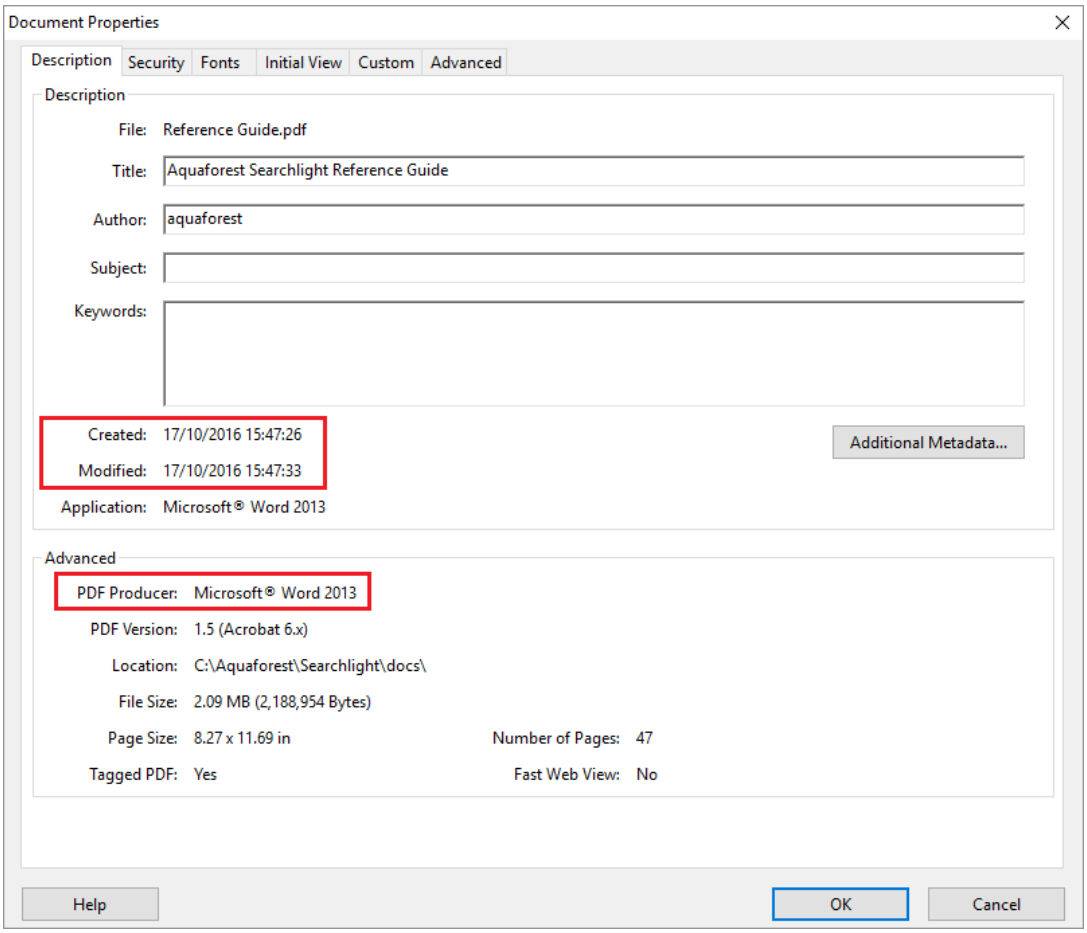
For this to be manipulated, the Searchlight service must be running with sufficient administrative privileges.

- ** SharePoint metadata vs. PDF metadata

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



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.

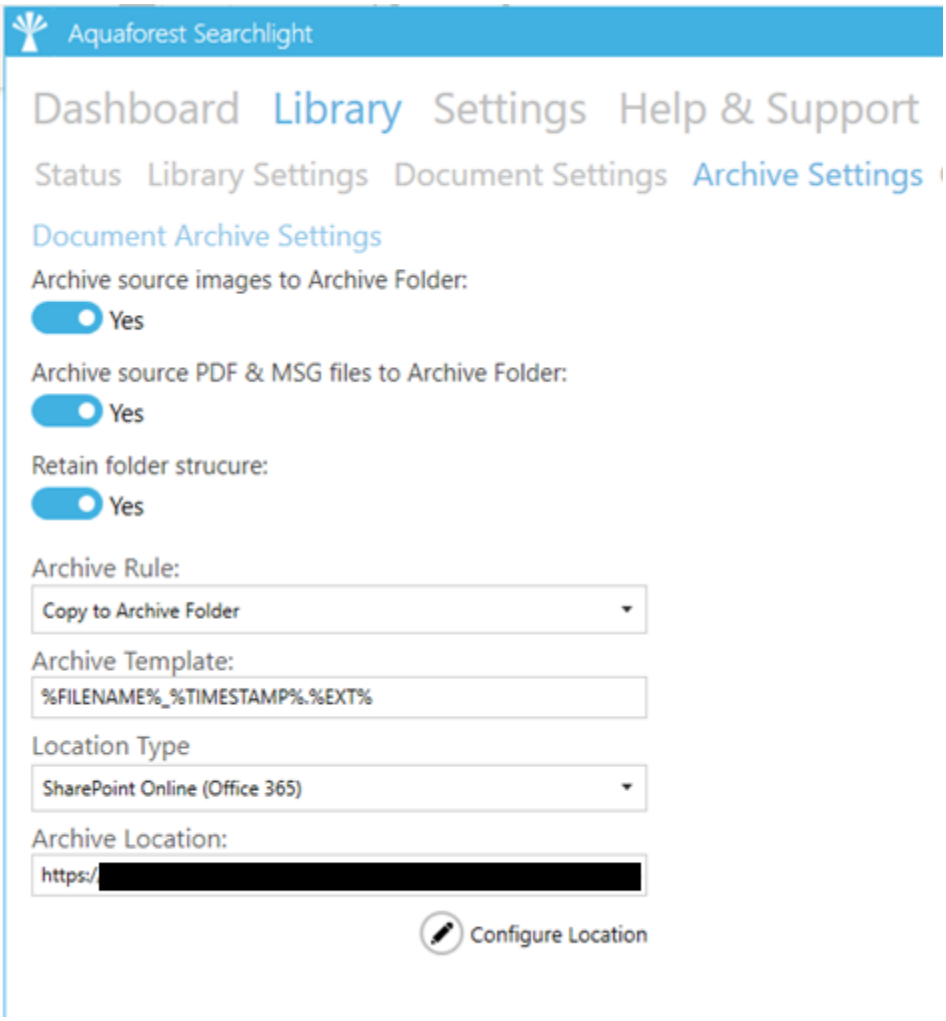
SharePoint Settings		Searchlight Settings	Created Date retained ?	Created By retained ?	Modified Date retained ?	Modified By retained ?
Create Major Versions	Create Minor Versions	Publish Major Version				
×	×	n/a*	✓	✓	✓	✓
✓	×	n/a*	✓	✓	✓	✓
✓	✓	×	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓	×

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				
×	n/a	✓	✓	✓	✓
×	n/a	✓	✓	✓	✓
✓	n/a	✓	✓	✓	✓
✓	n/a	✓	✓	✓	✓

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.
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.

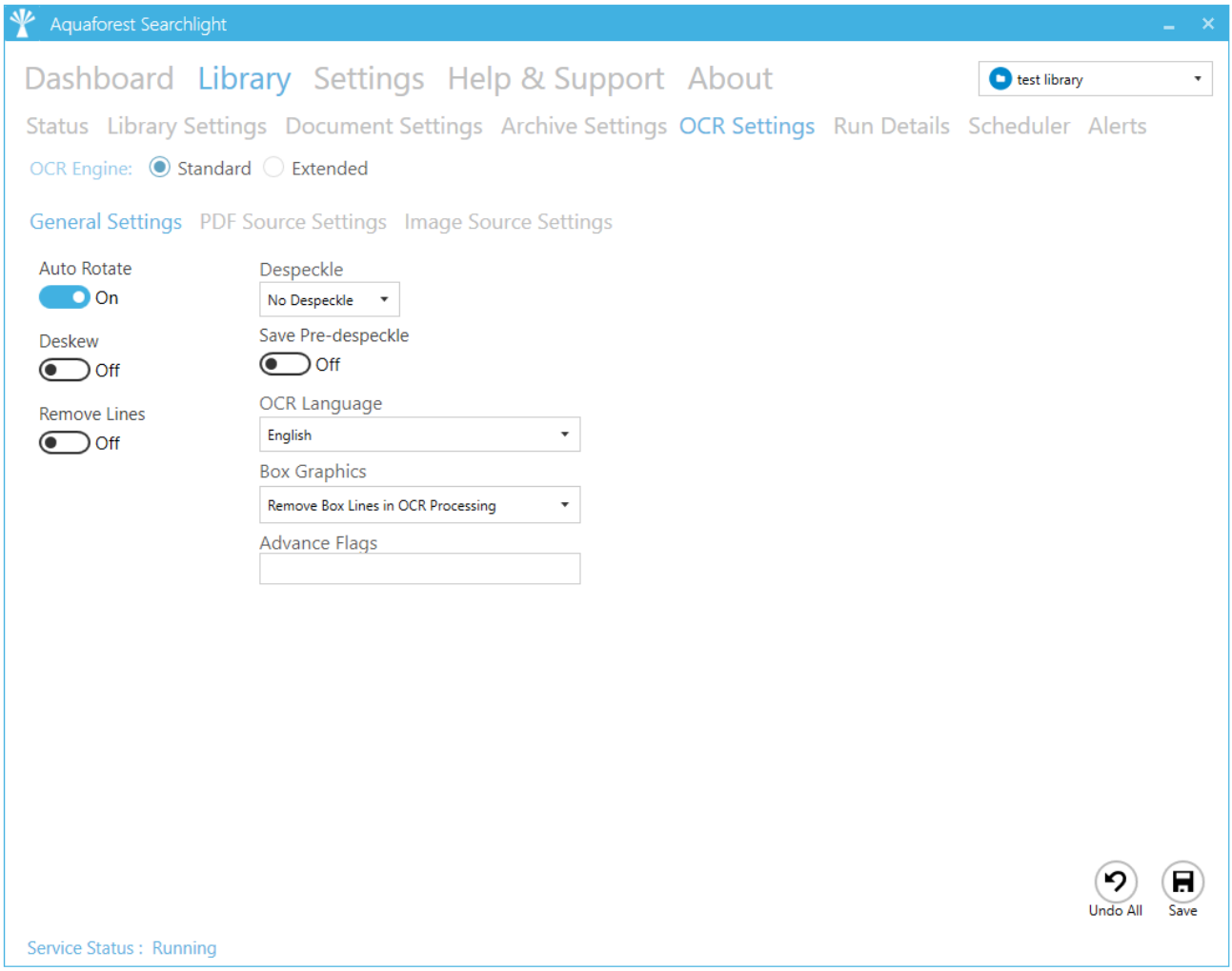
- Aquaforest engine: "[installation path]\ocr\Properties.xml"
- Extended (IRIS) engine: "[installation path]\extendedocr\Properties.xml"

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

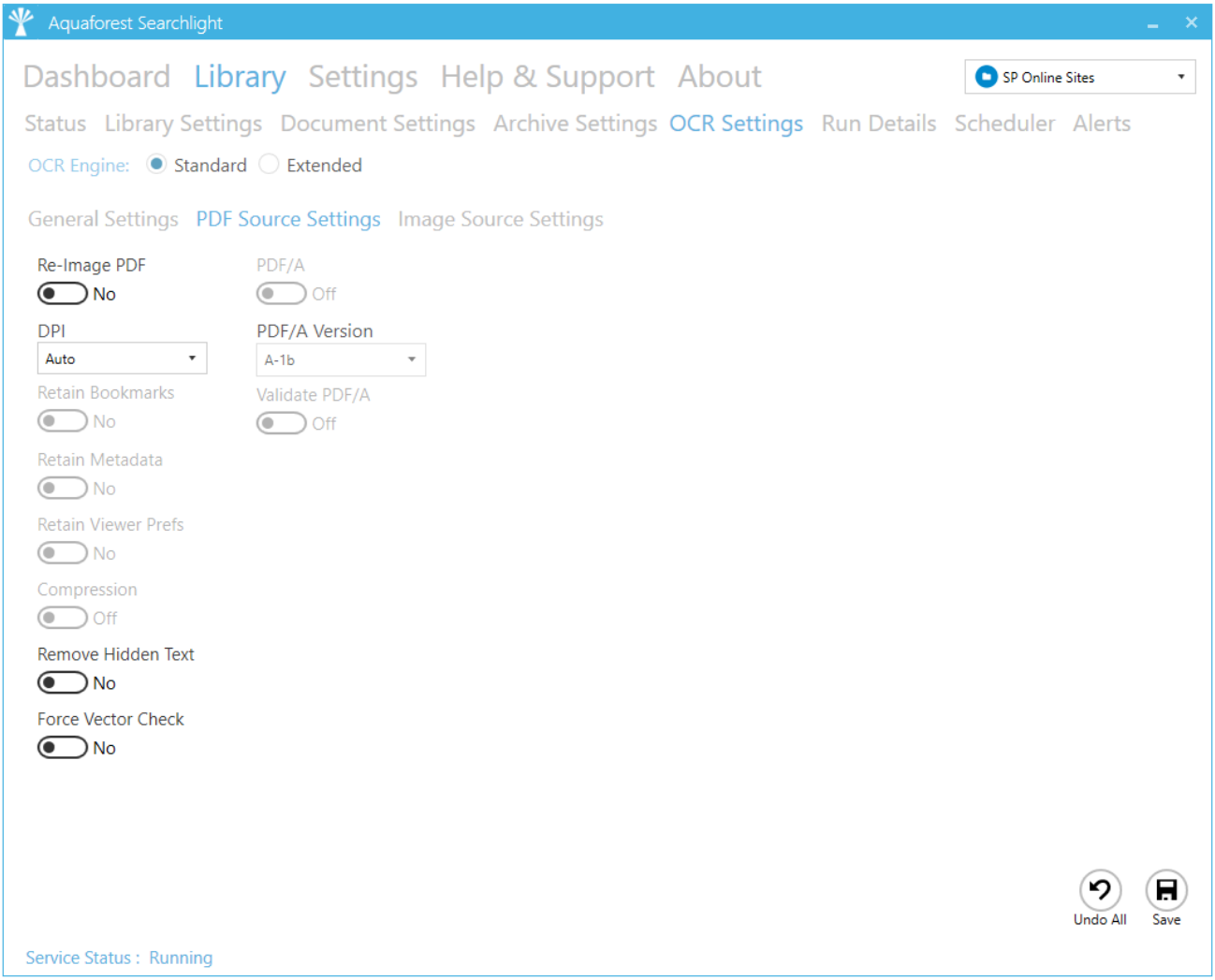
6.3.5.1.1 General Settings



Setting	Description
General Settings	
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	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. To ensure that the OCR engine can be forced to process such areas there are two options: *"Treat all Graphics Areas as Text".* This option will ensure the entire document is processed as text. *"Remove Box Lines in OCR Processing".* 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).
Advanced Flags	Command line flags to be passed through to the underlying executable. Contact support@aquaforest.com for details on using this field.

6.3.5.1.2 PDF Source Settings



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

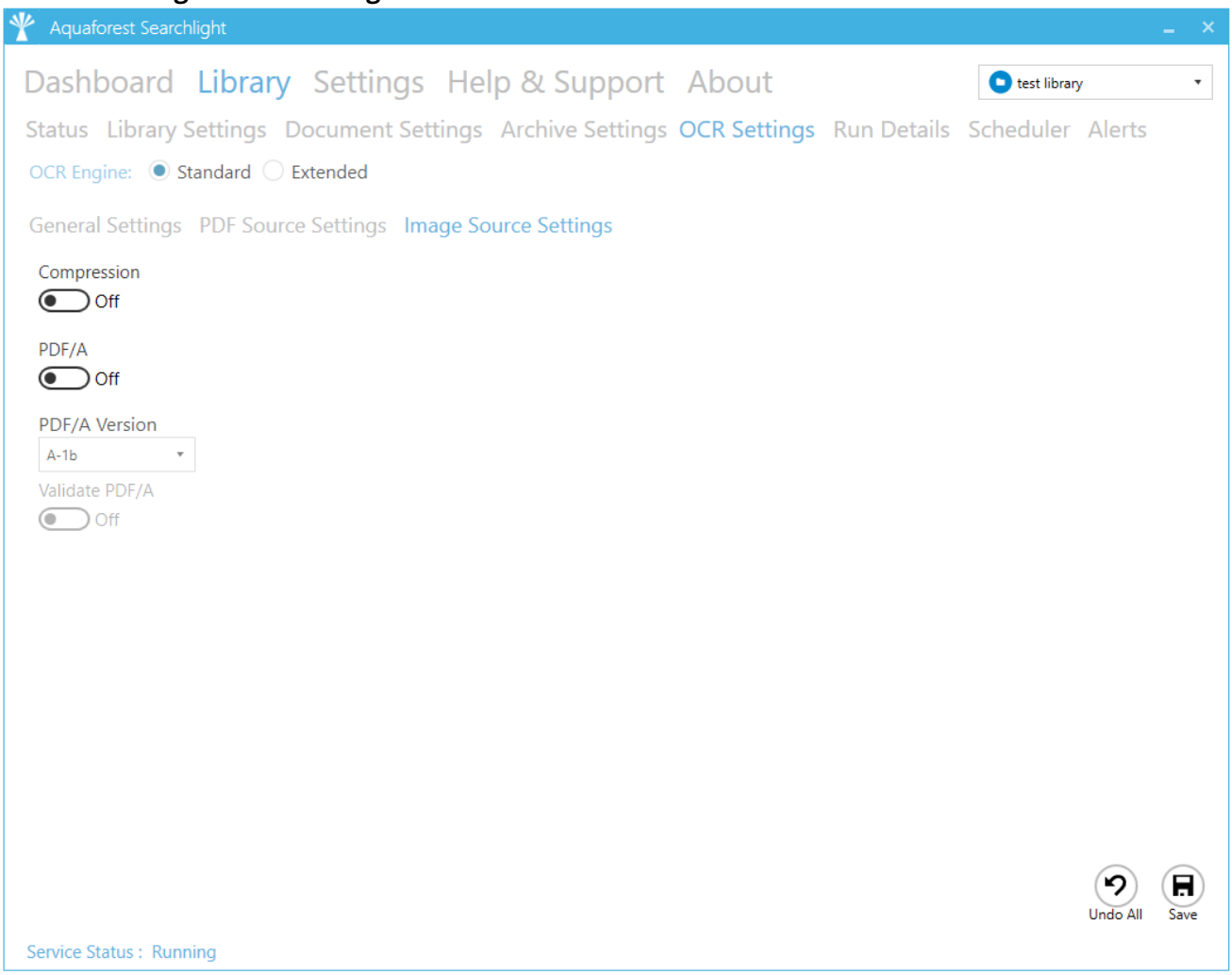
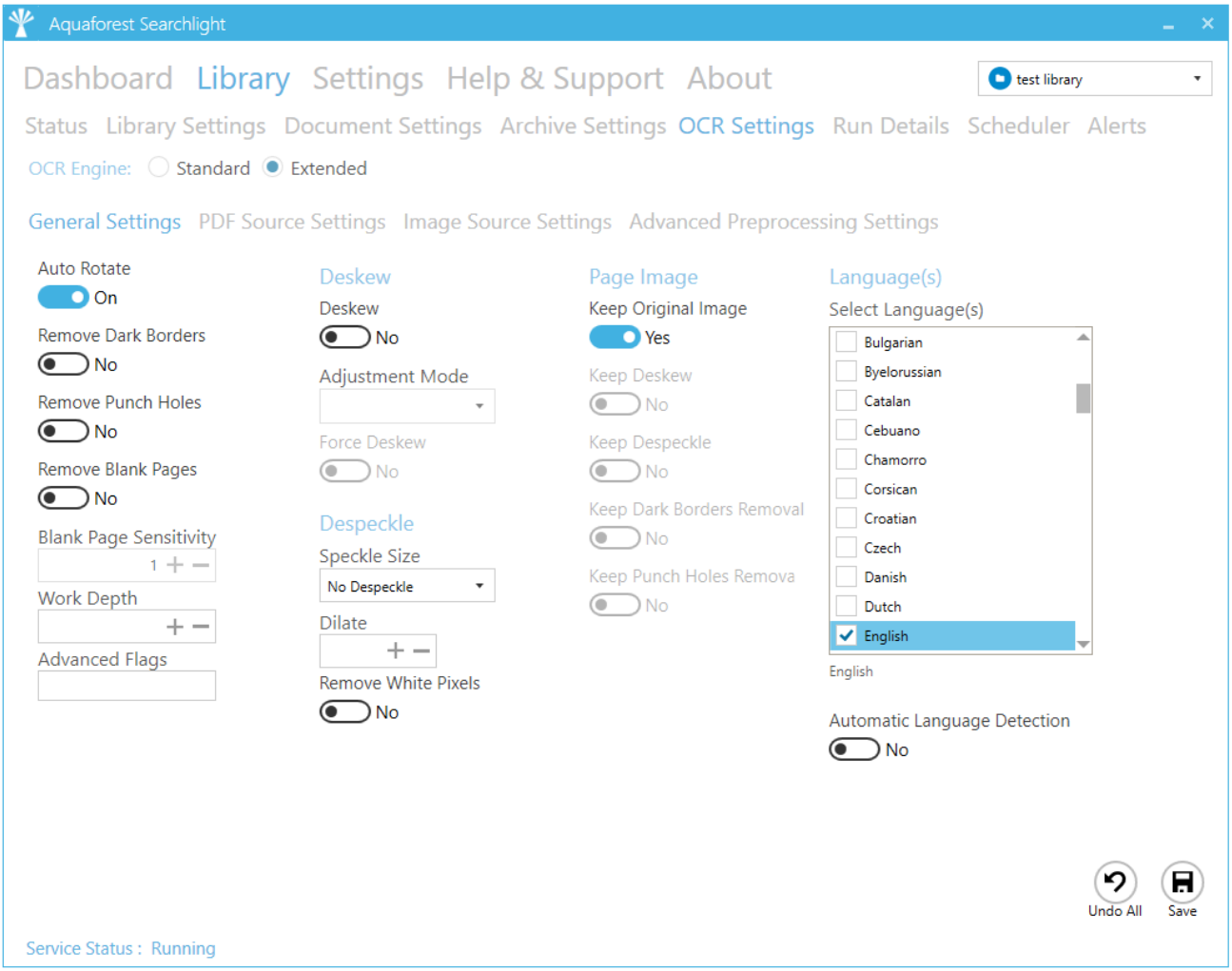


Image Source Settings	
Compression	The image(s) in the output PDF file will be compressed using JBIG2 (for black and white image) or MRC (for color images) which can dramatically reduce the output size of PDFs.
PDF/A	Switch on to make sure the output PDF conforms to the PDF/A standards.

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

6.3.5.2 Extended OCR Settings

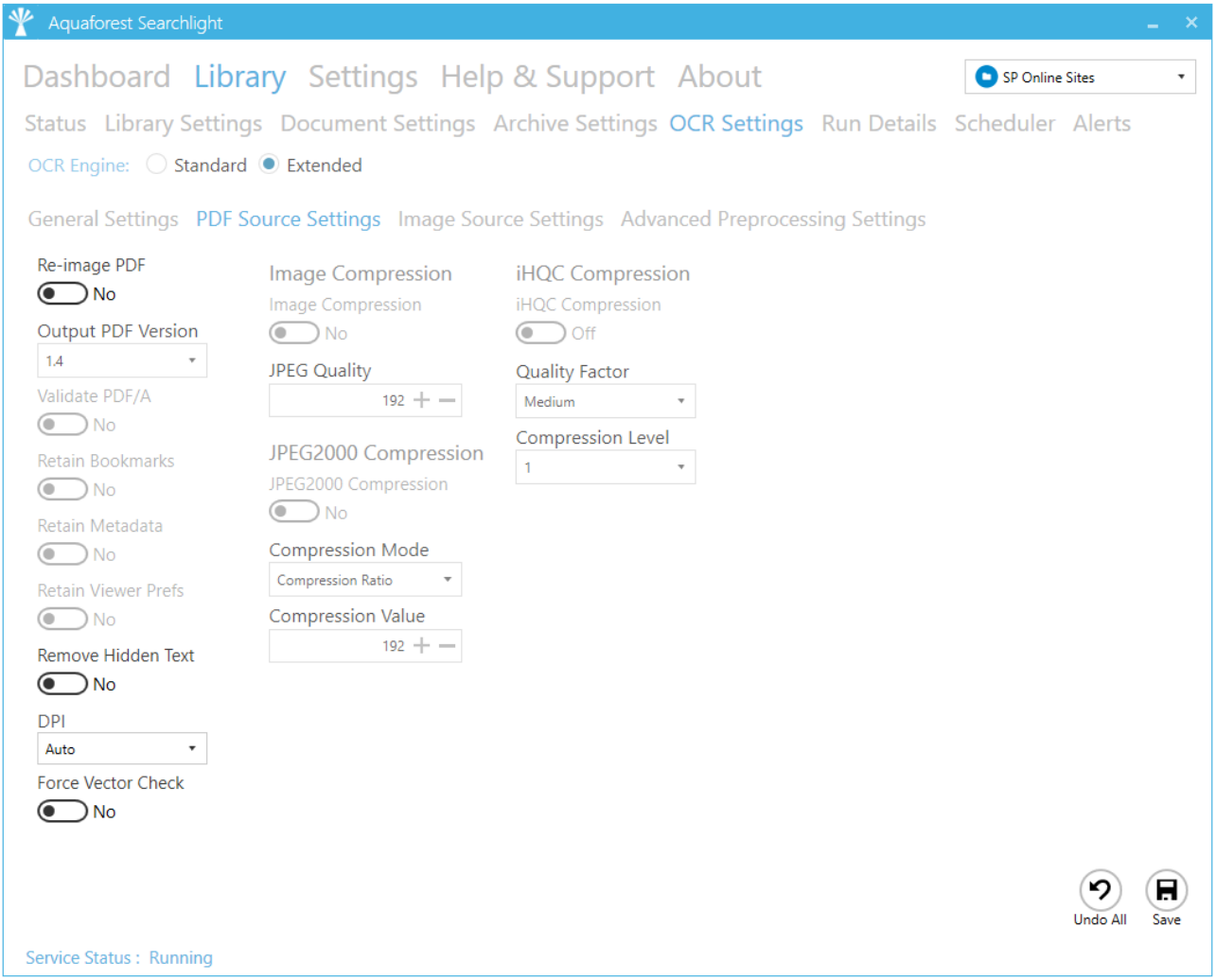
6.3.5.2.1 General Settings



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 selected pre-processing has been 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.

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. **Note:** • Only a maximum of 8 languages can be selected. • Only the English language can be used in conjunction with an Asian language

6.3.5.2.2 PDF Source Settings



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'.
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

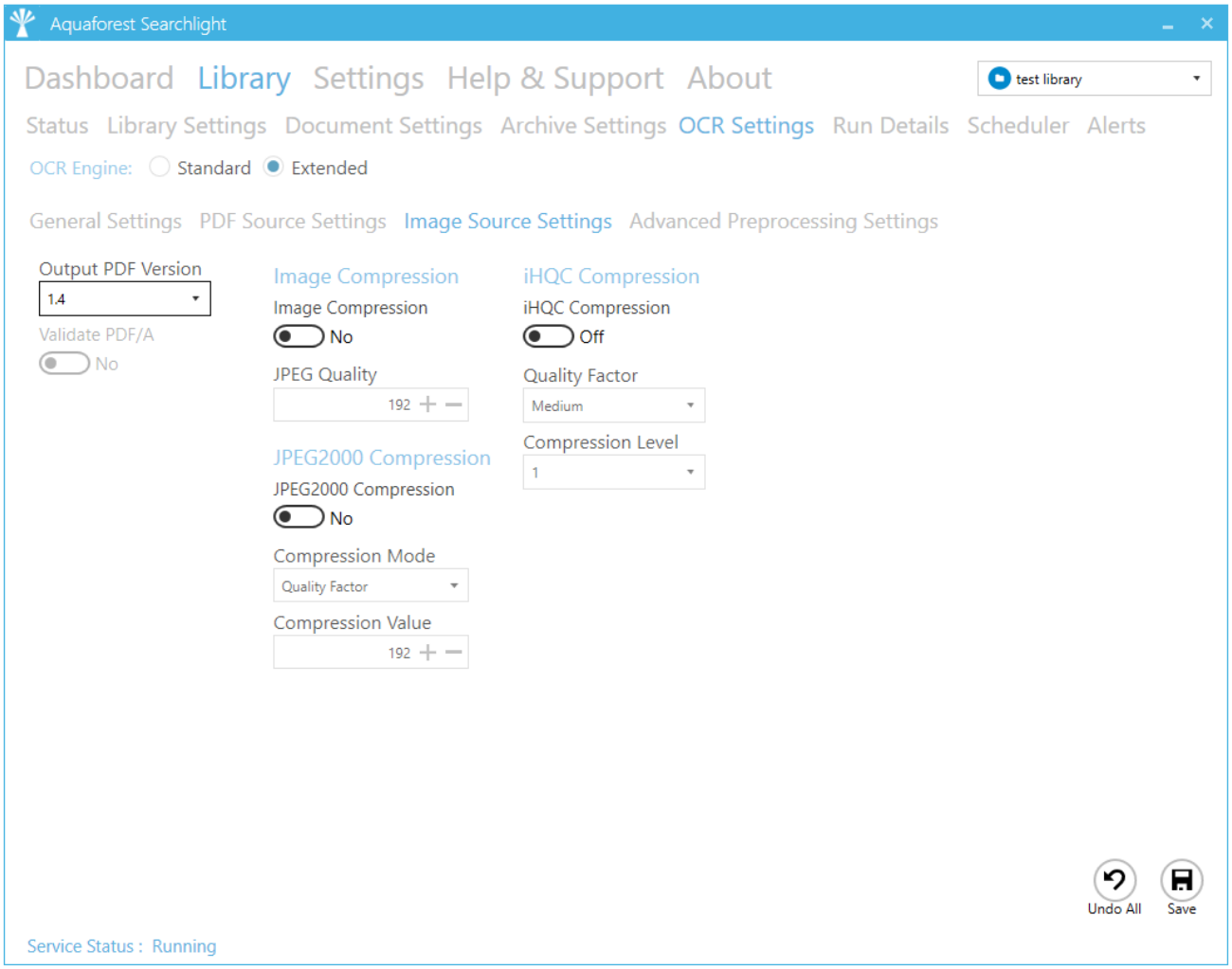
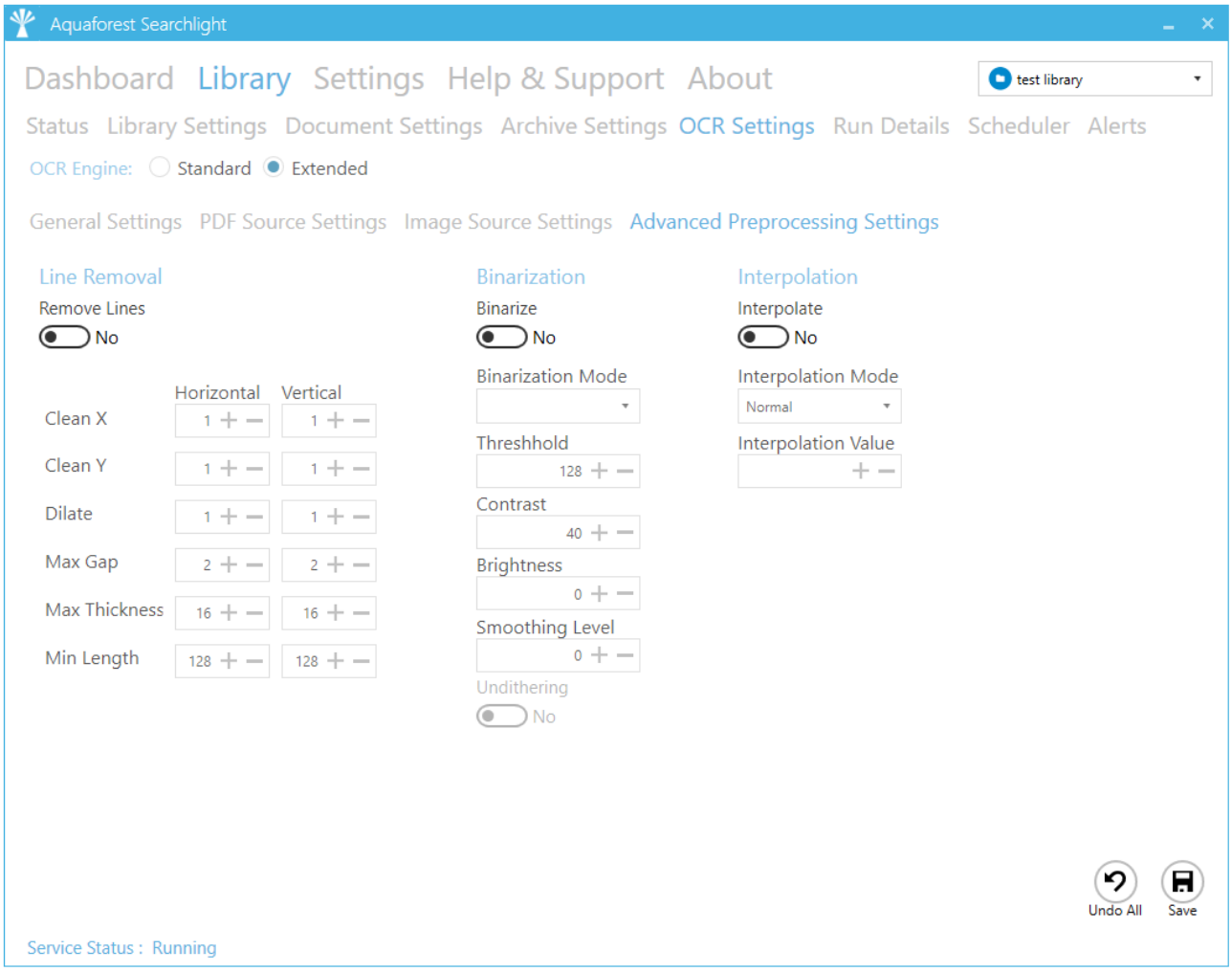


Image Source Settings	
Output PDF Version	This determines the PDF version of the generated PDF.
Image Compression	Compress color JPEG images in generated PDFs
JPEG Quality	This parameter (0 – 255) determines the compression/quality of color JPEG images in generated PDFs. 0 gives the smallest file size 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.

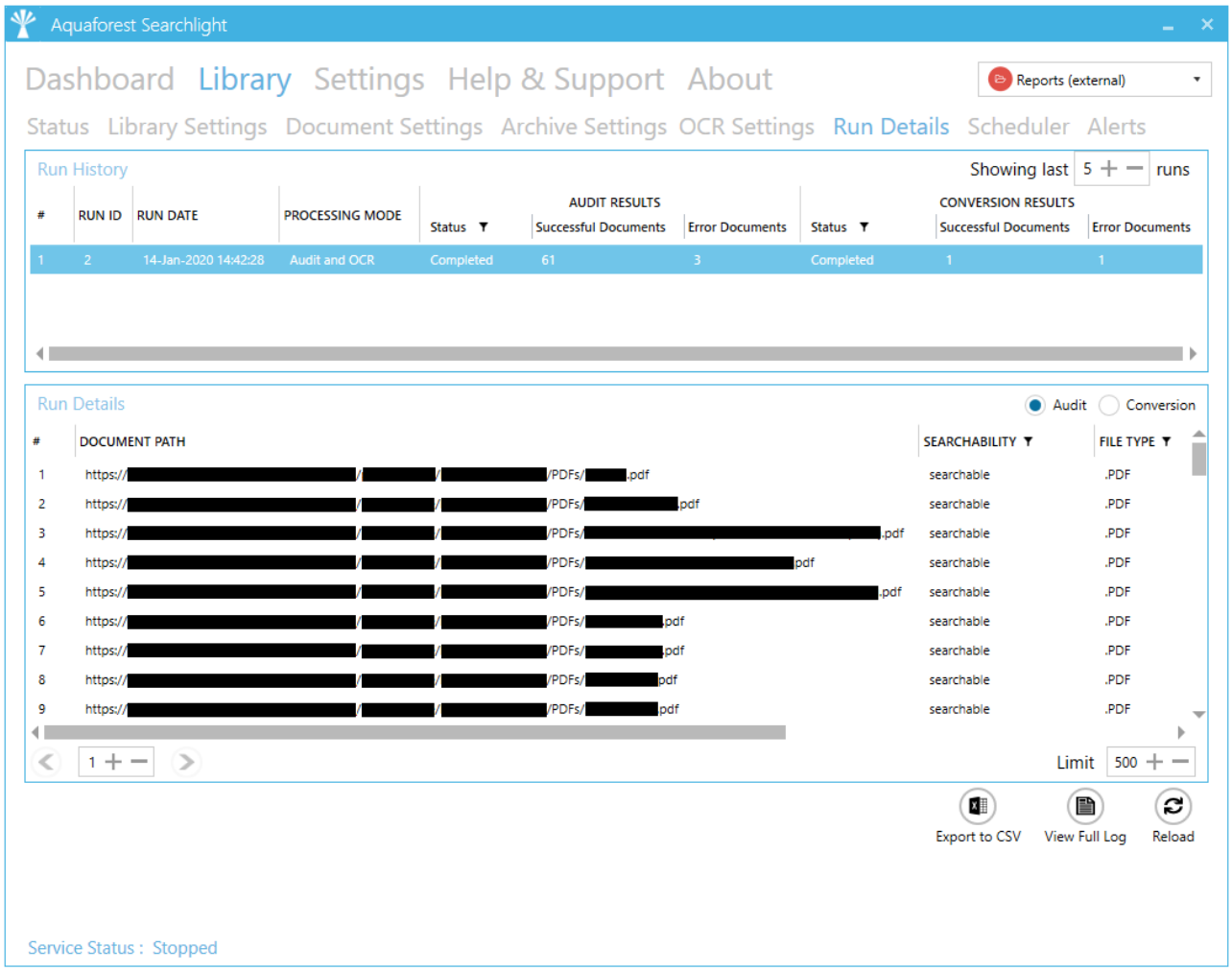
6.3.5.2.4 Advanced Pre-processing Settings



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



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.

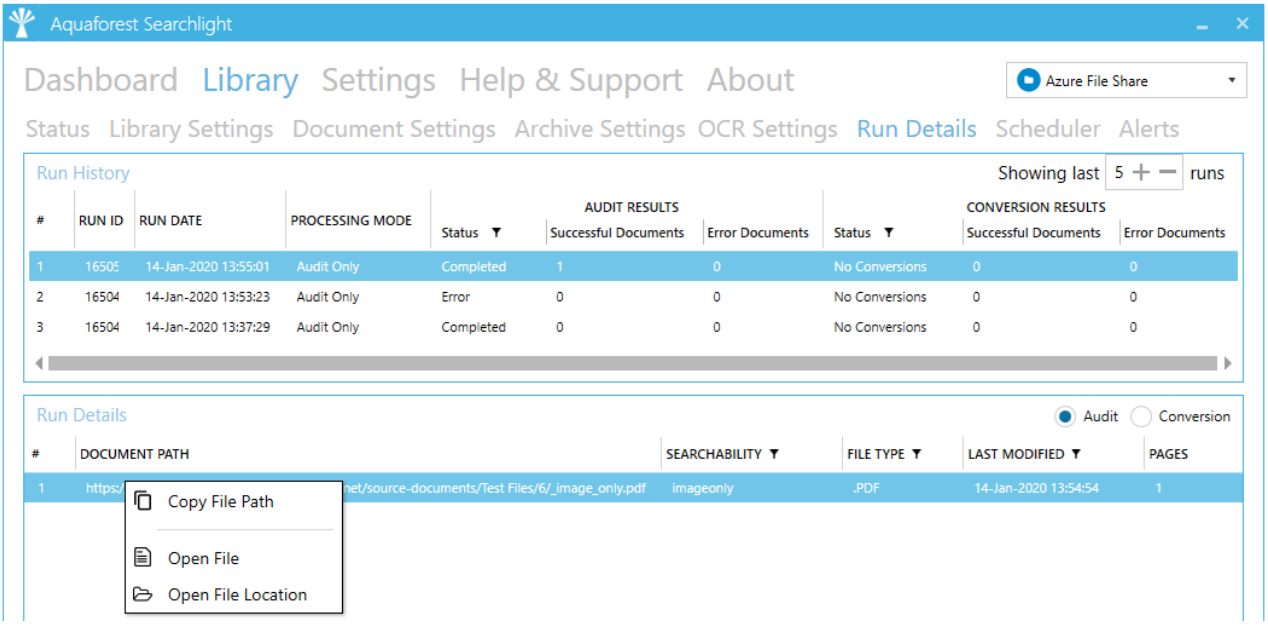
Use **Export to CSV** to export the run details to CSV file.

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

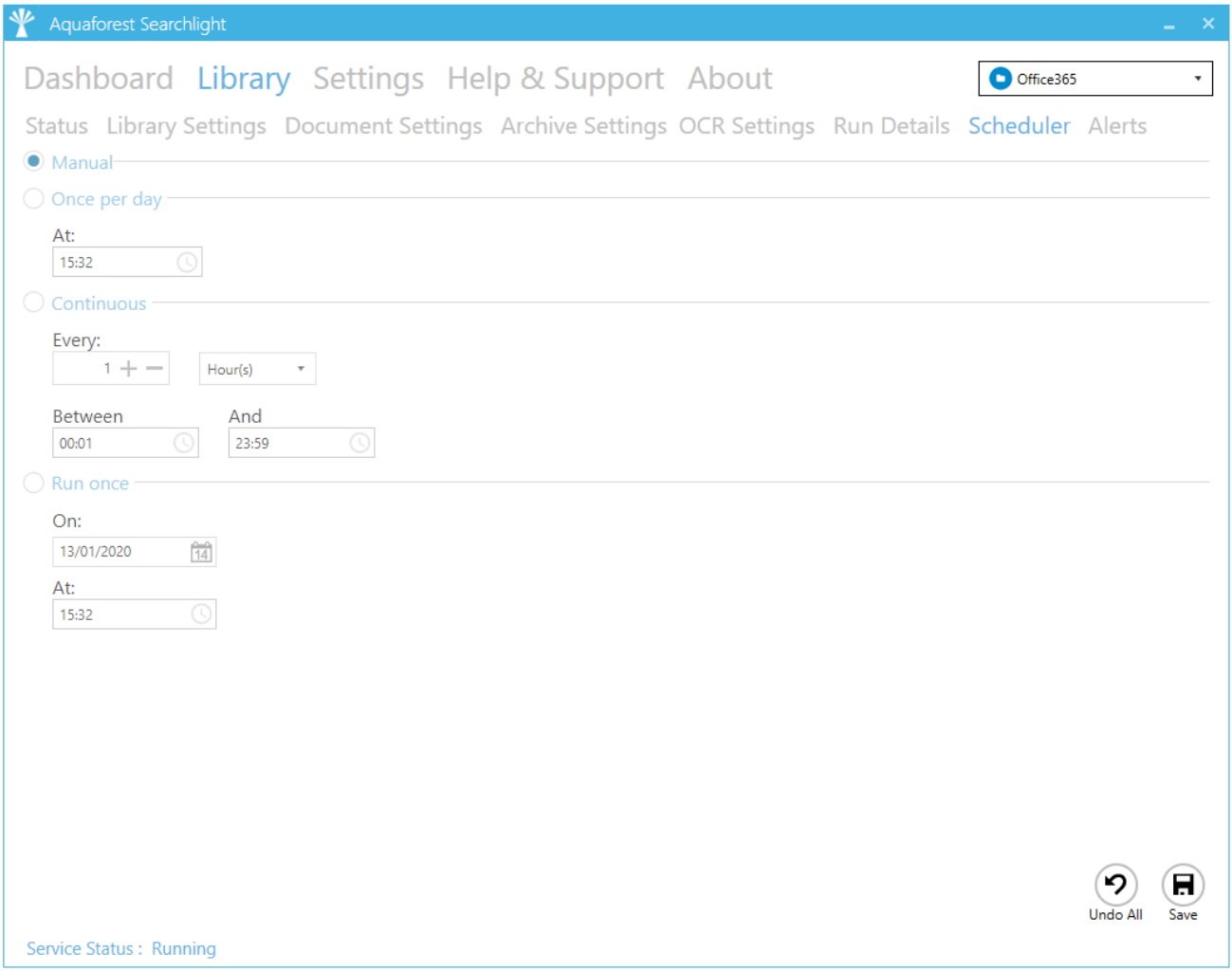
6.3.6.1 Run Details Context Menu

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)



6.3.7 Scheduler Settings

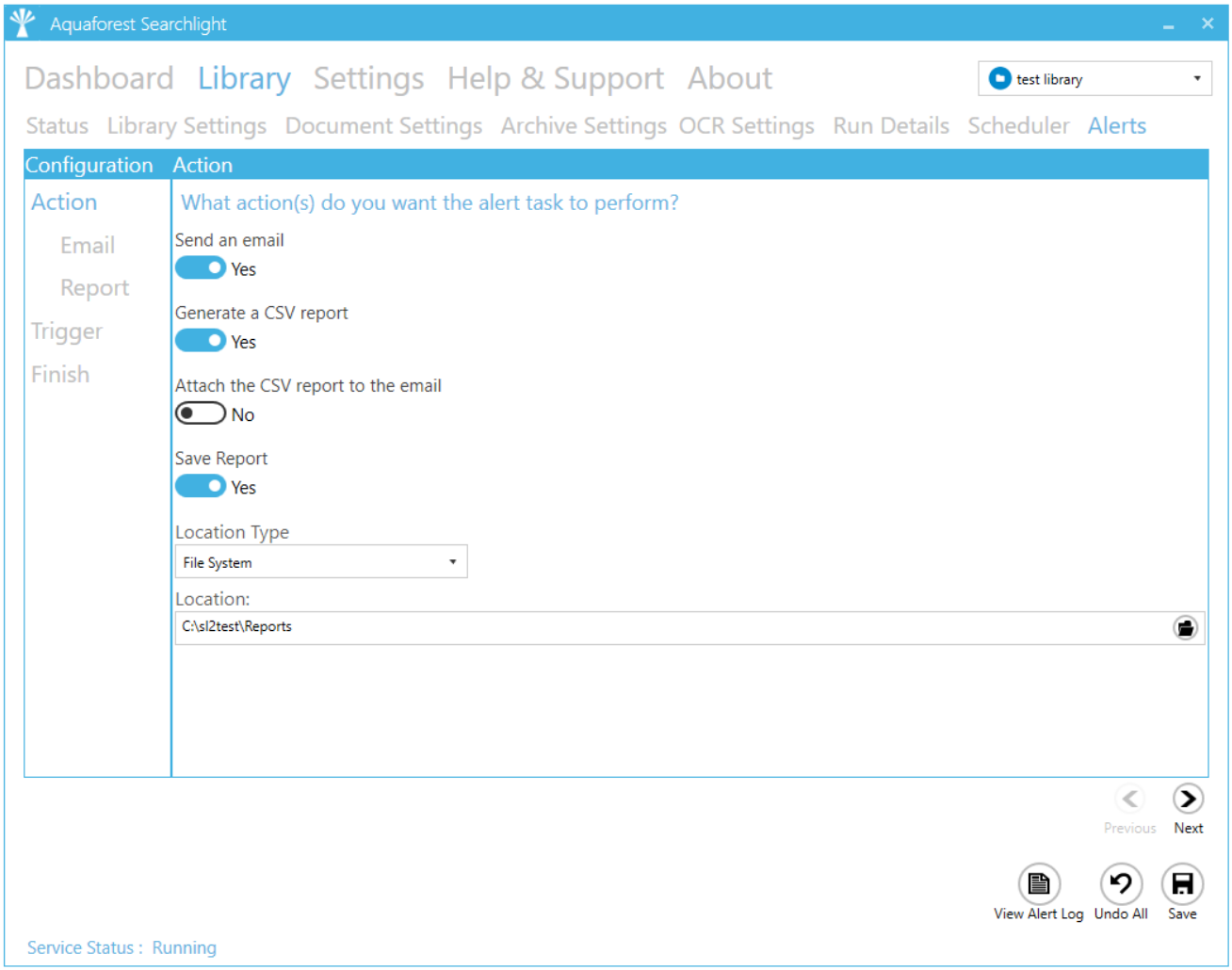


Setting	Description
Manual	This means that the document library must be run manually by clicking on the "Run" button on the dashboard.

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

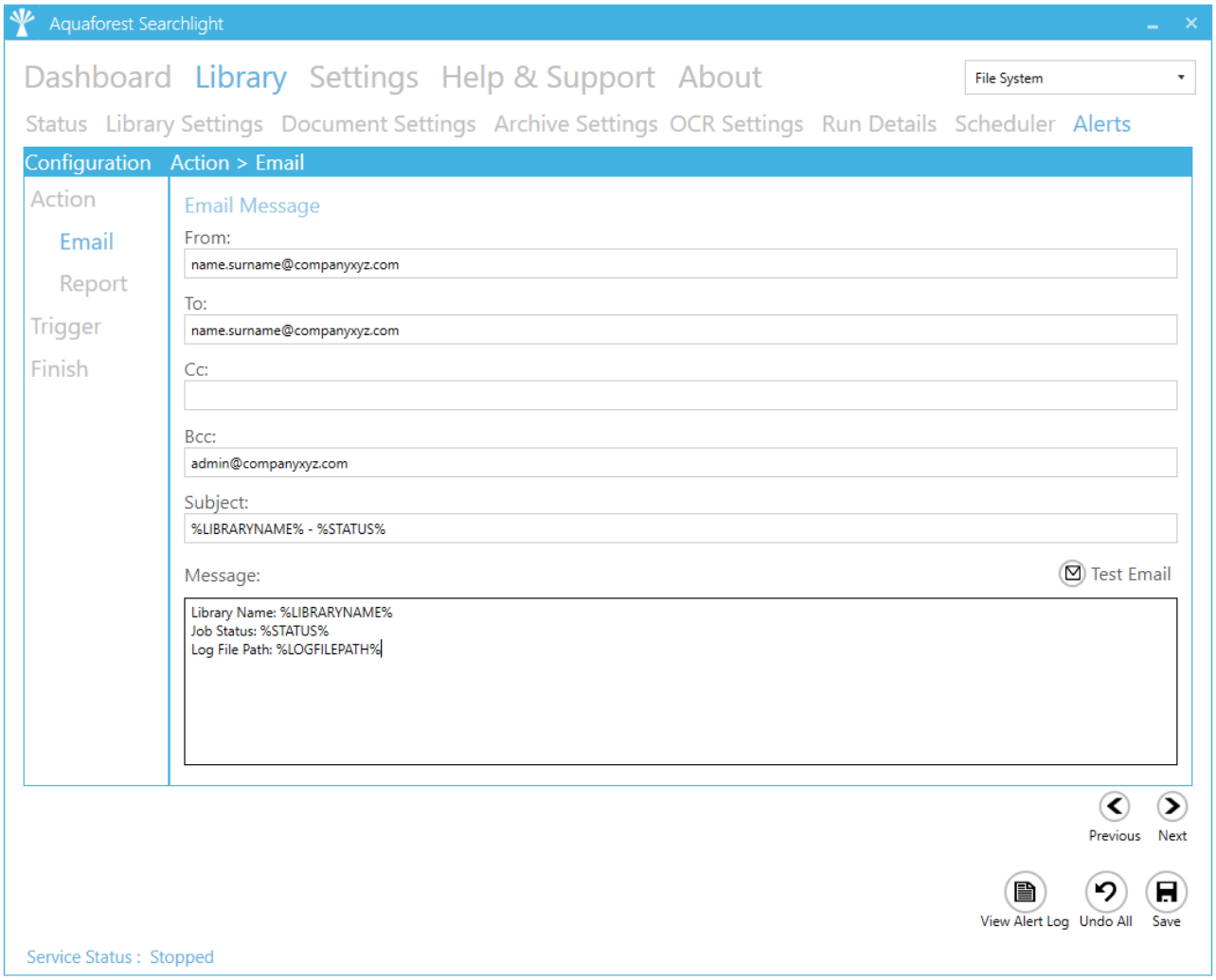
6.3.8.1 Action



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

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

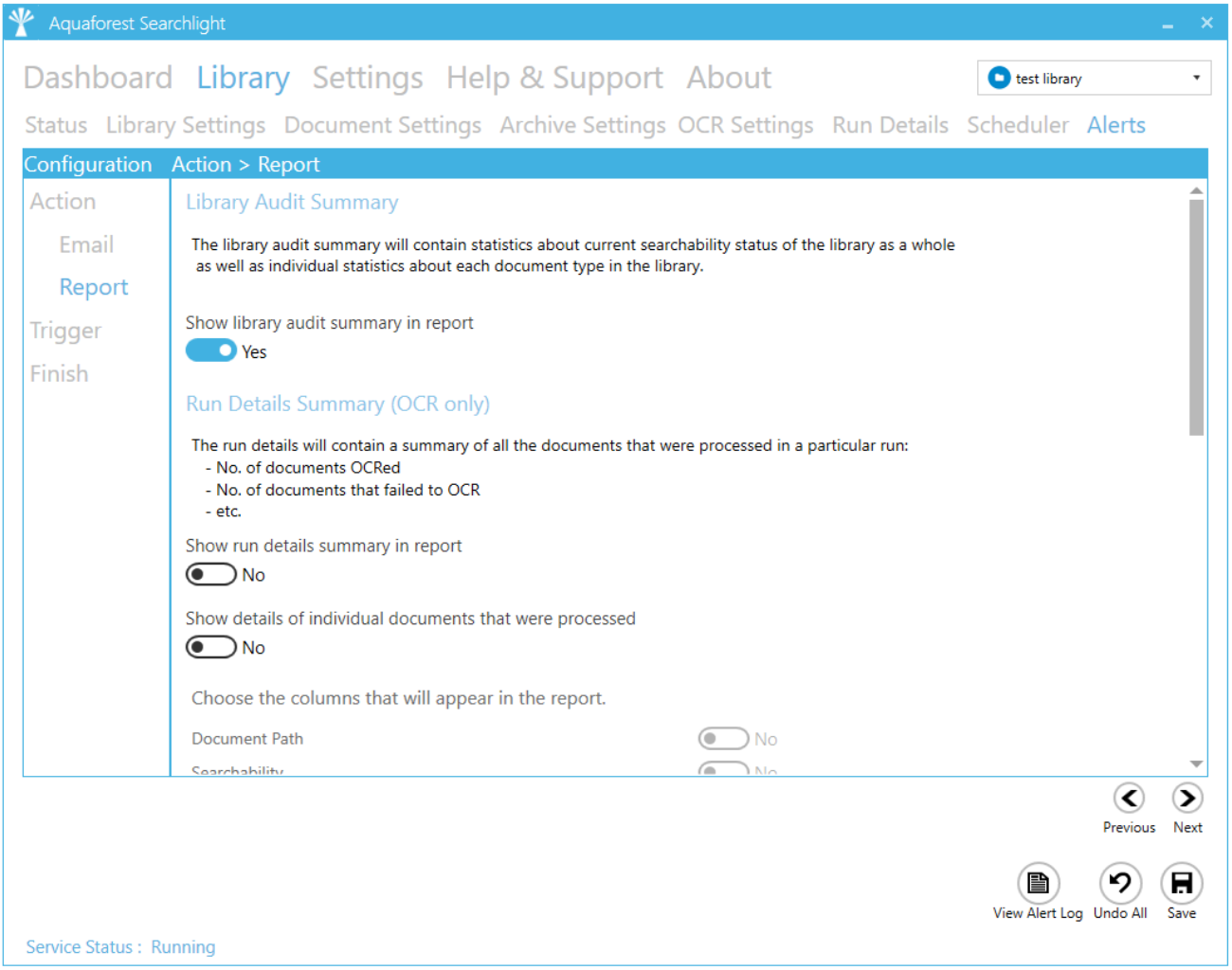
6.3.8.2 Email



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	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
Message	The email message to send. You can use the following templates within the email message: • **%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 for the library • **%ERRORMESSAGE%** - will be replaced by any error messages that occurred during the library run

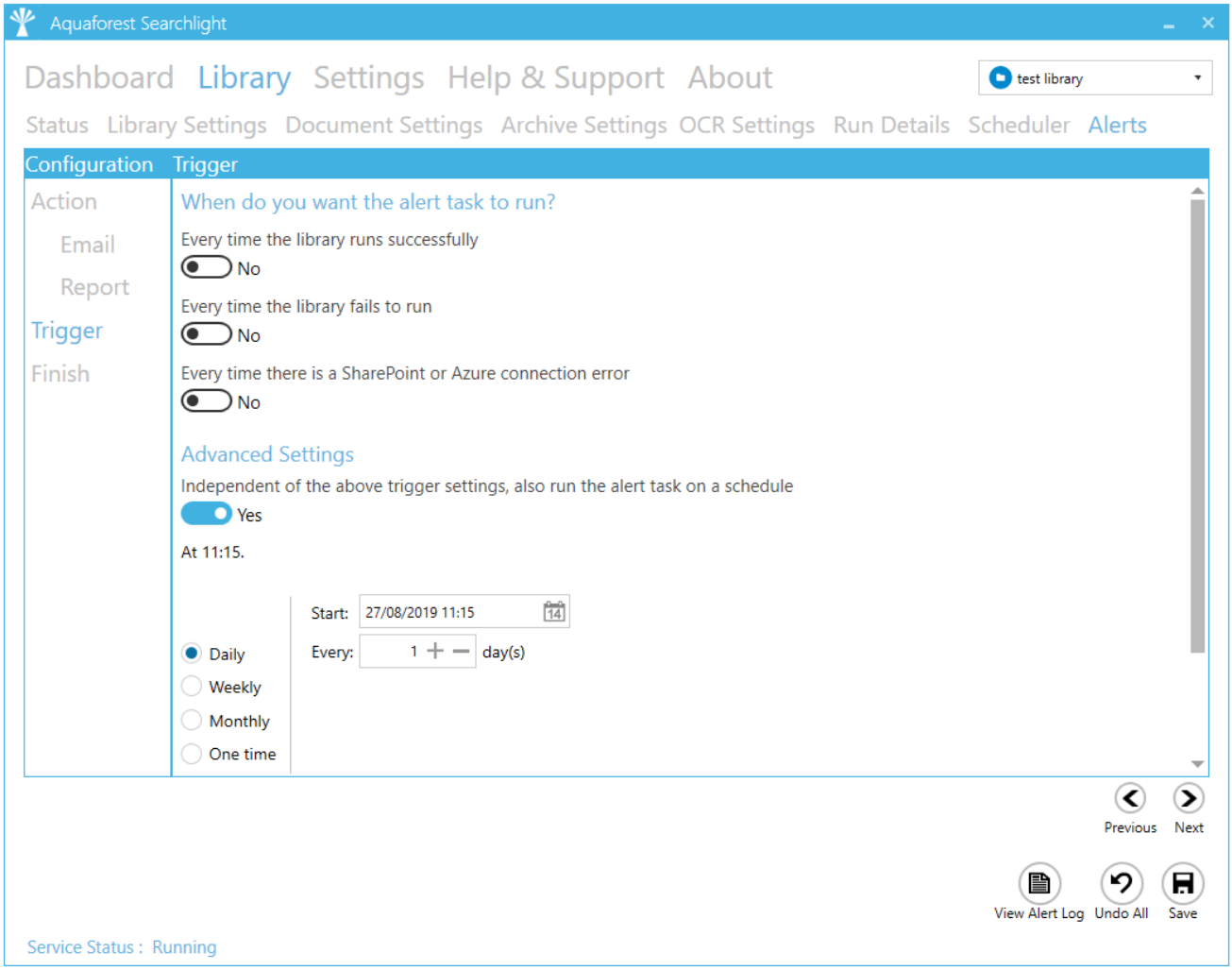
6.3.8.3 Report



Report

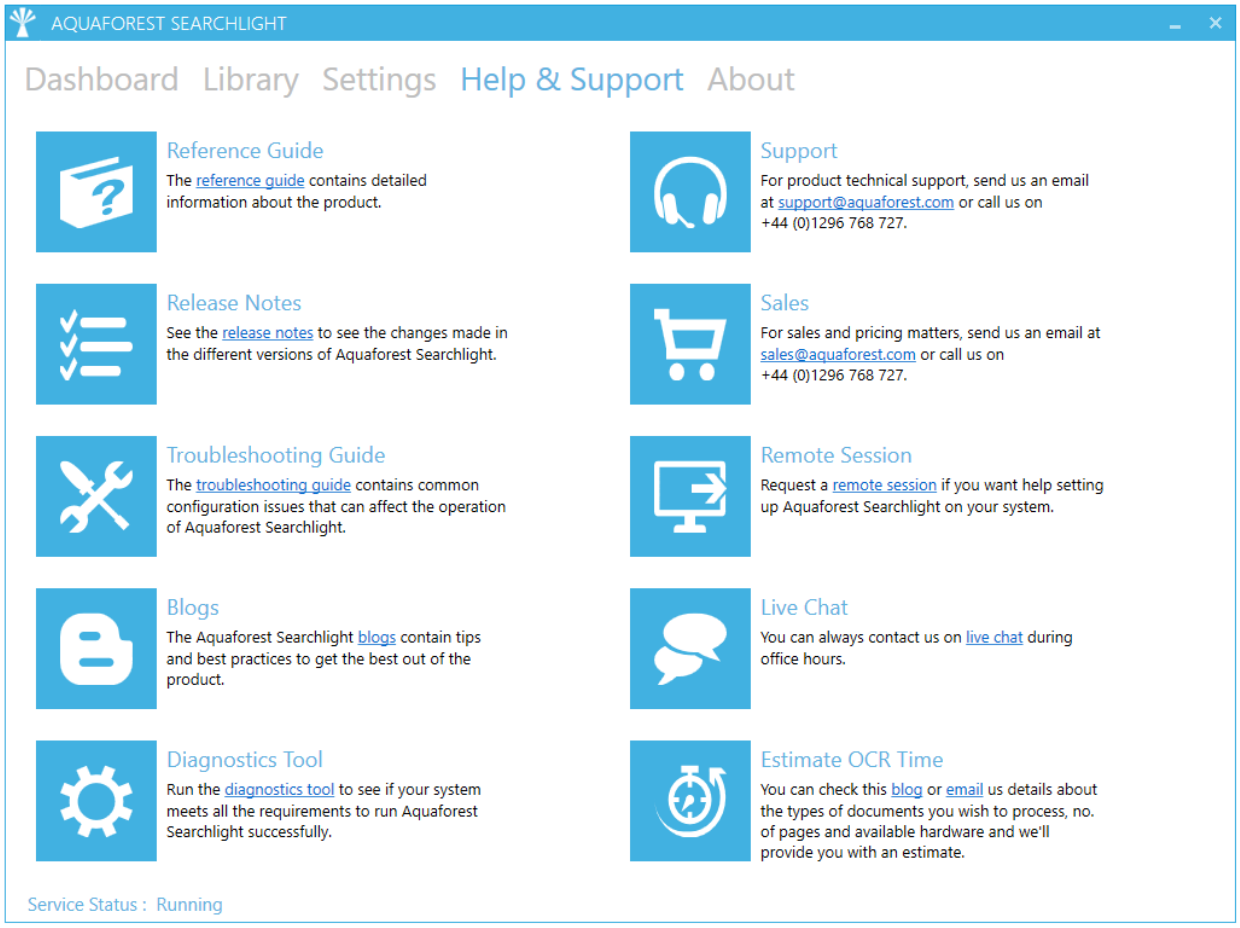
Show library audit summary in report	The library audit summary contains statistics about the current searchability status of the library as a whole, as well as individual statistics about each document type in the library.
Run Details Summary (OCR only)	
Show run details summary in report.	The run details summary lists all the documents that were processed in a particular run including: • Number of documents OCRed • Number of documents that failed to OCR
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



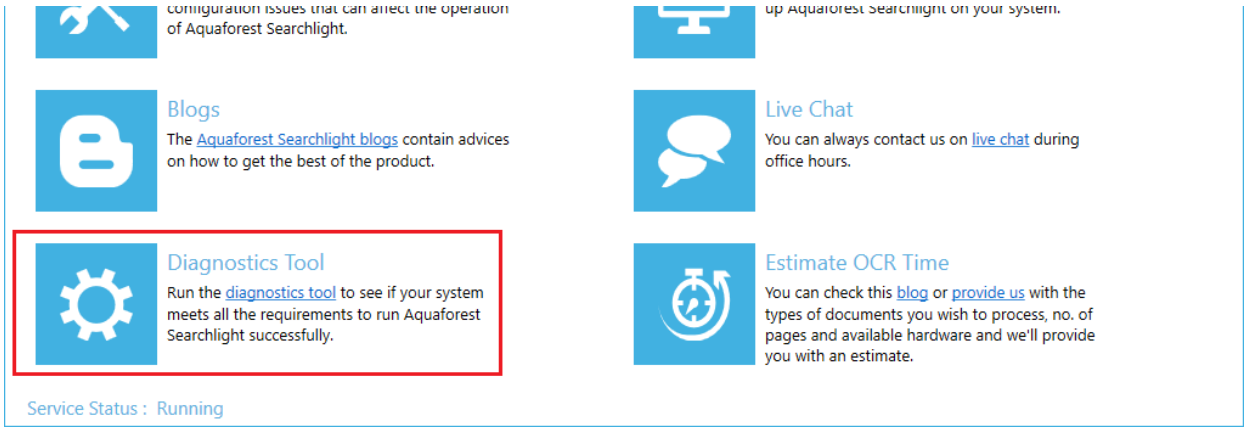
Trigger	
Alert is triggered	Every time the library runs successfully. Every time the library fails to run. Every time there is a SharePoint or Azure connection error
Advanced Settings	Independent of the above trigger settings, the alert can be scheduled to run daily, weekly (on selected days), monthly 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 for further investigation.

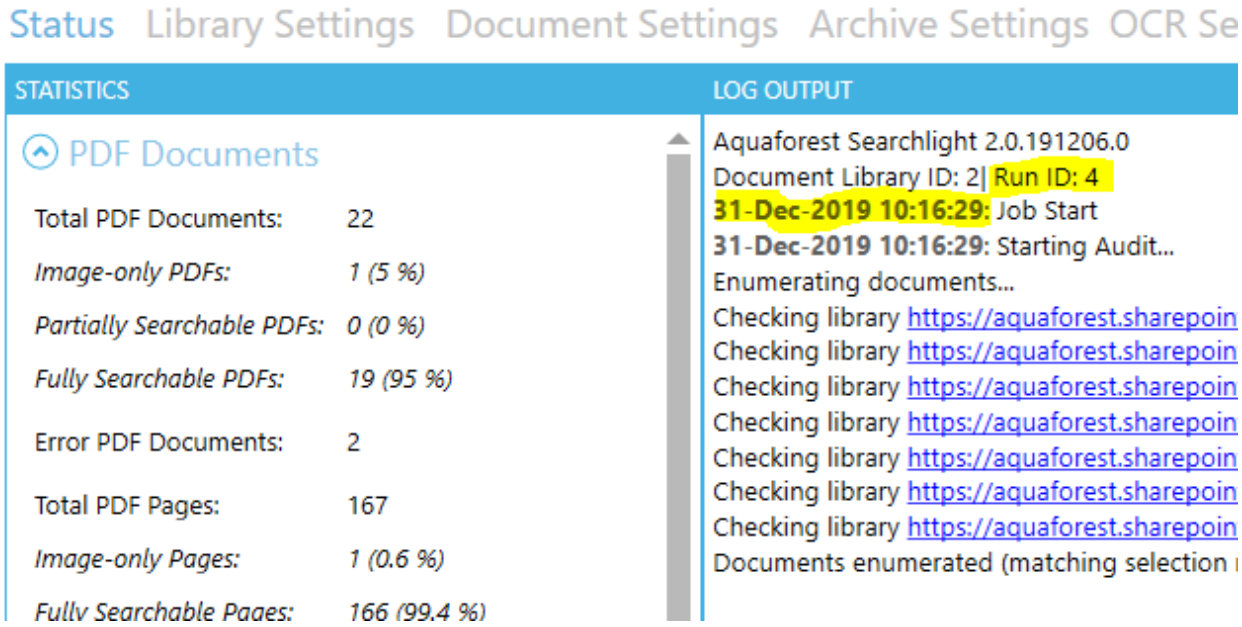
6.4.2 Database Clean-up Tool

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.

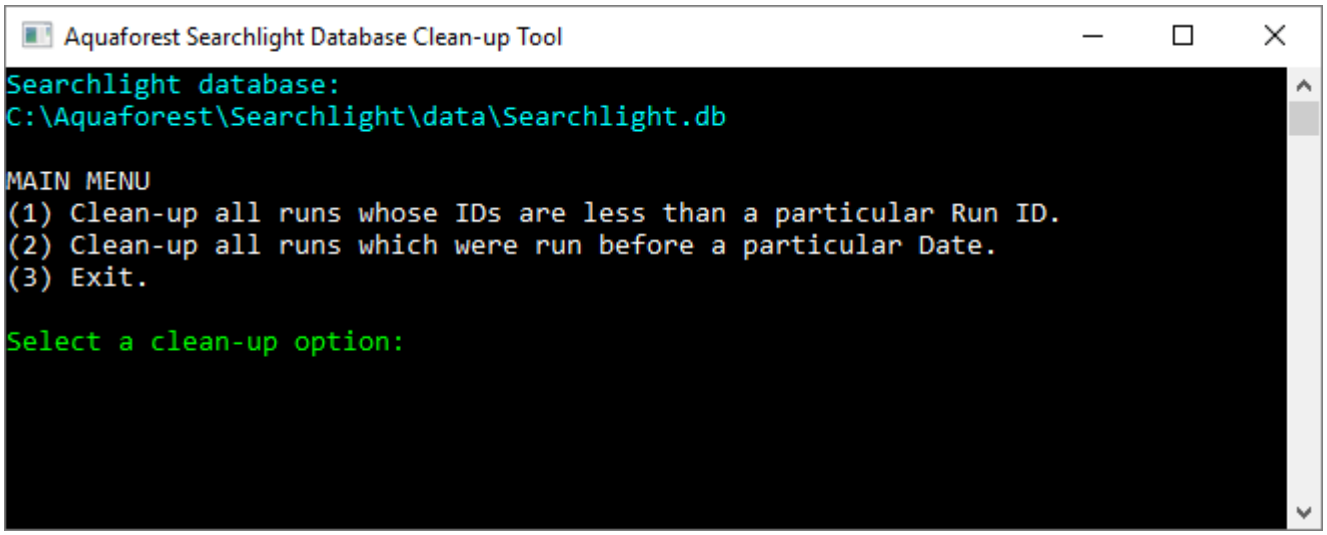
Searchlight comes with a command line tool that will try to compact the database by deleting logs from previous runs.

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.

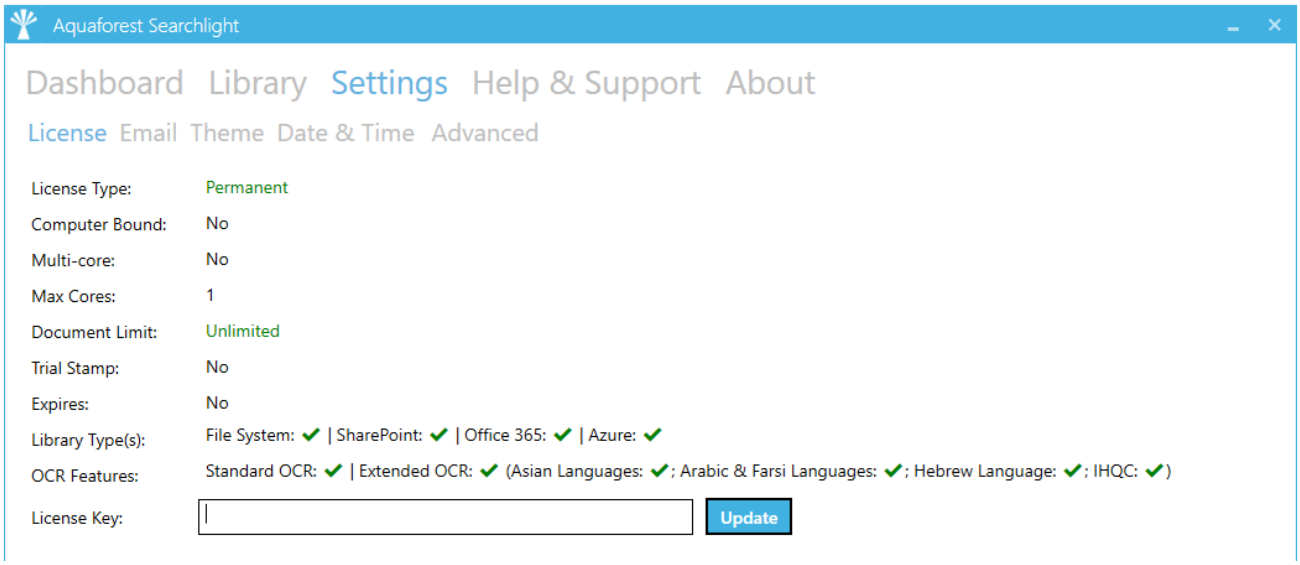


With this information, log in as an Administrator and start the command line tool.



6.5 Settings

6.5.1 License Settings

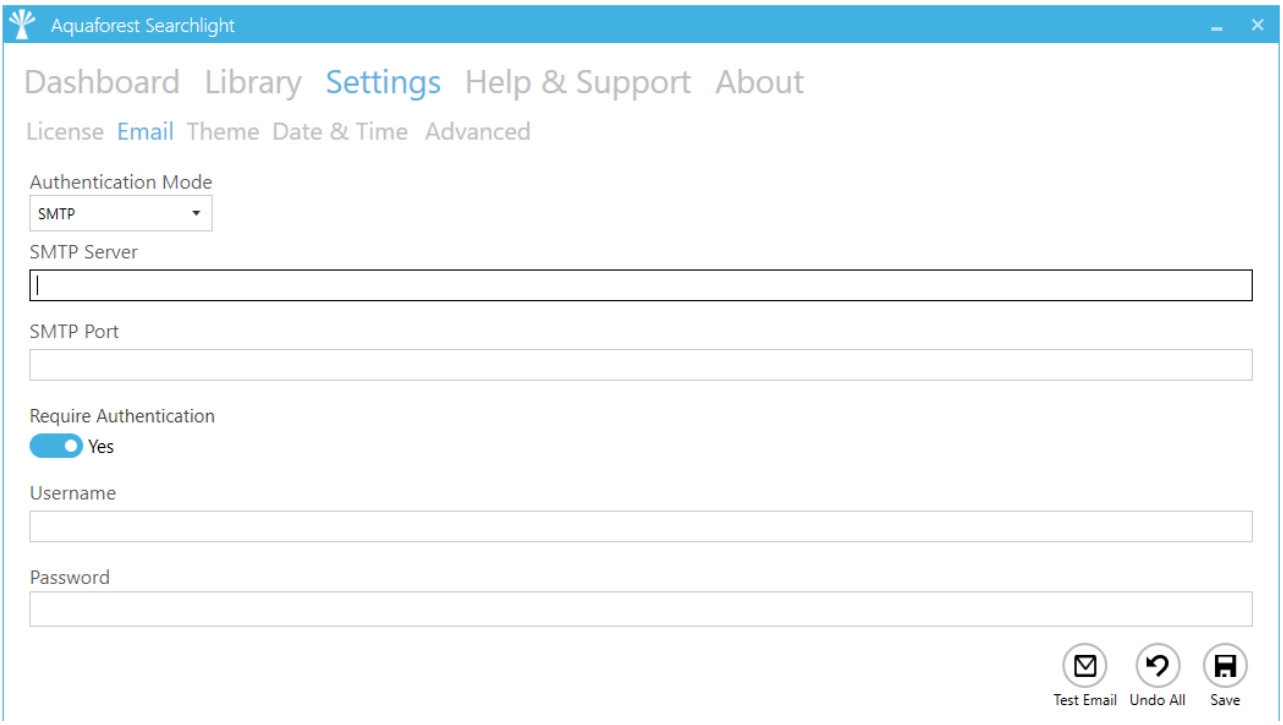


Setting	Description
License Type	Trial or Permanent
Computer Bound	Whether the license is computer bound or not computer bound
Computer Identifier	The name of the computer if the license is computer bound
Multi-core	Whether or not the license allows the use of multiple cores for processing
Max Cores	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.
Trial Stamp	Whether or not the OCRed 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.

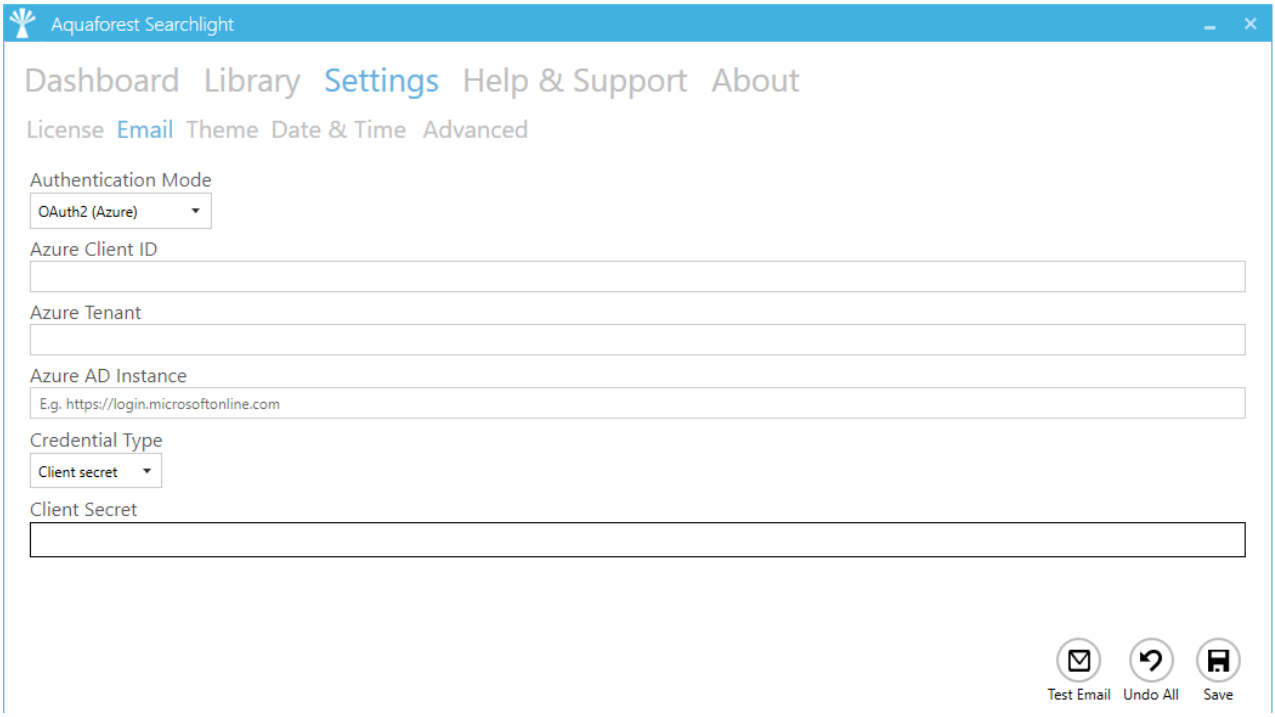
6.5.2.1 SMTP



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

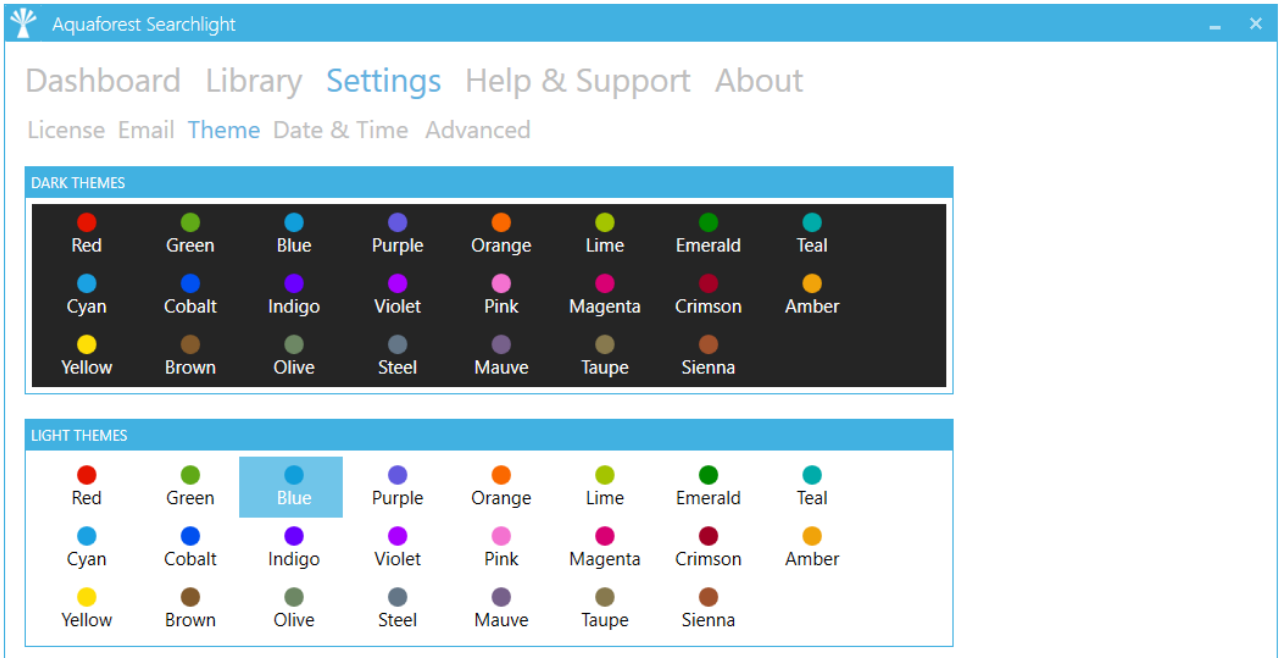
For additional details on OAuth2 authentication, refer to "**Exchange Online OAuth2 Configuration.pdf**" document in the **docs** folder where Searchlight is installed.



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: https://login.microsoftonline.com
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

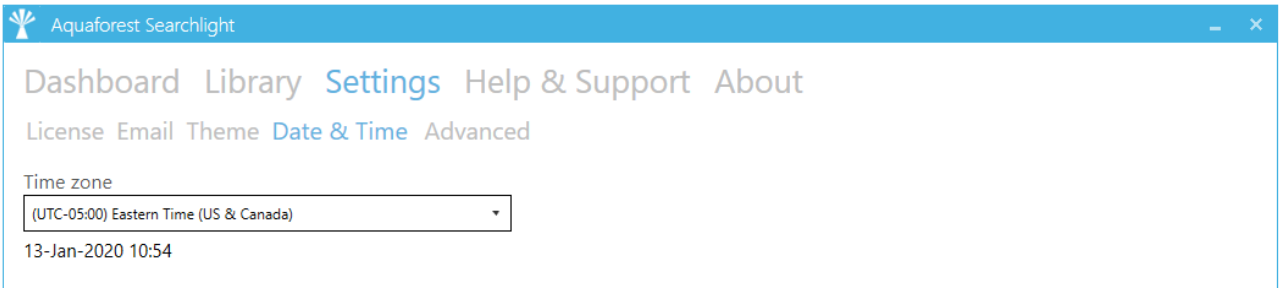
6.5.3 Themes

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



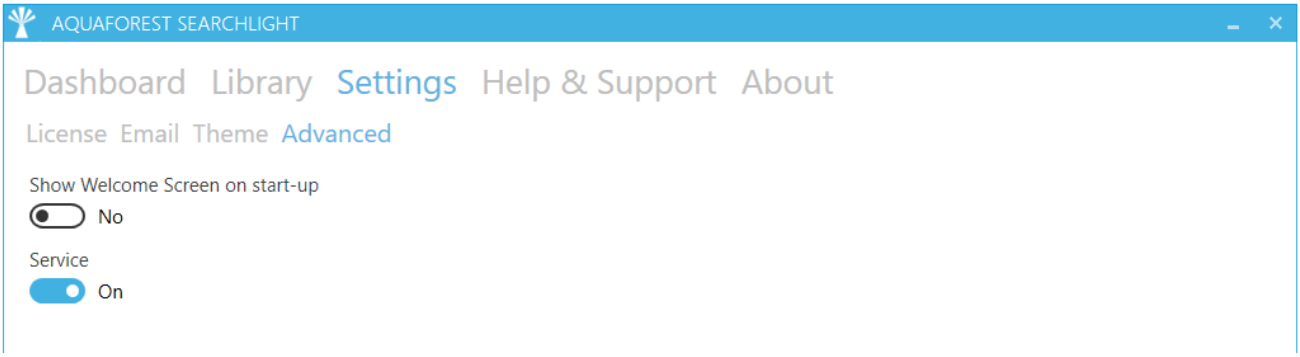
6.5.4 Date & Time

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 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.

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	Searchlight may re-OCR documents that have already been processed previously if its modified date in SharePoint has changed since the last time it was processed and process "Fully Searchable" and/or "Partially Searchable" options are set in the Document Settings. The modified date can change if a document is replaced by a new one or its metadata/properties are modified in SharePoint. To avoid re-processing these documents again irrespective of whether the modified has changed, set this setting to “true”. The default value is false.
sharePointFailCheckinComment	When a SharePoint document is successfully OCRed, 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. To force Searchlight to add a comment to the non-OCRed document in SharePoint, specify a comment in this setting.

Setting	Description
failOnPixelLimit	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. Extended OCR has the following image limits: • Max Height = 32,768 pixels • Max Width = 32,768 pixels • Max Size = 75,000,000 pixels
pdfTextOperators	The PDF text operators that need to be present in a page to consider it searchable.
downloadAndUploadRetries sharePointRequestRetries	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: • downloadAndUploadRetries - used when downloading and uploading documents fail. • sharePointRequestRetries - used when executing SharePoint queries fail. 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.

Setting	Description
databaseRetries	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. 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.

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