
Aquaforest Searchlight Release Notes



Version 1.22

April 2017

1 Version 1.22

1.1 Enhancements

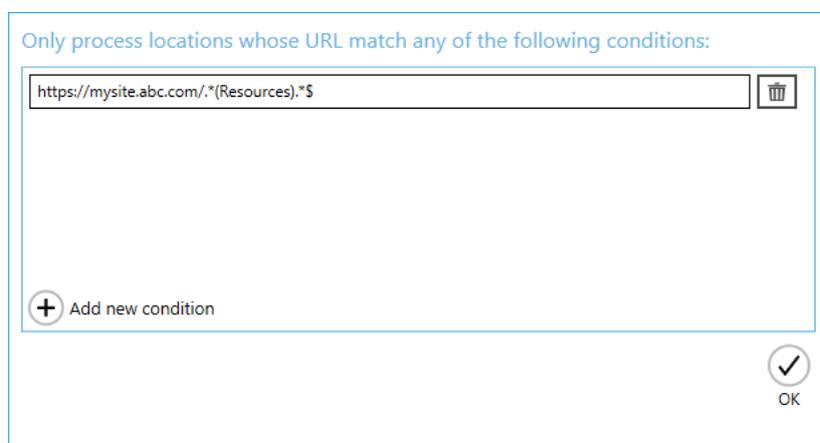
1.1.1 Advanced Pre-filtering

Added a new feature to pre-filter locations and documents by Regular Expressions before processing.

- Location filtering

This can be useful if you are processing a whole site collection but only want to include certain sites and libraries for processing.

For instance, you may want to only process sites and libraries containing the word "Resources" in their URL:



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



 Add new condition

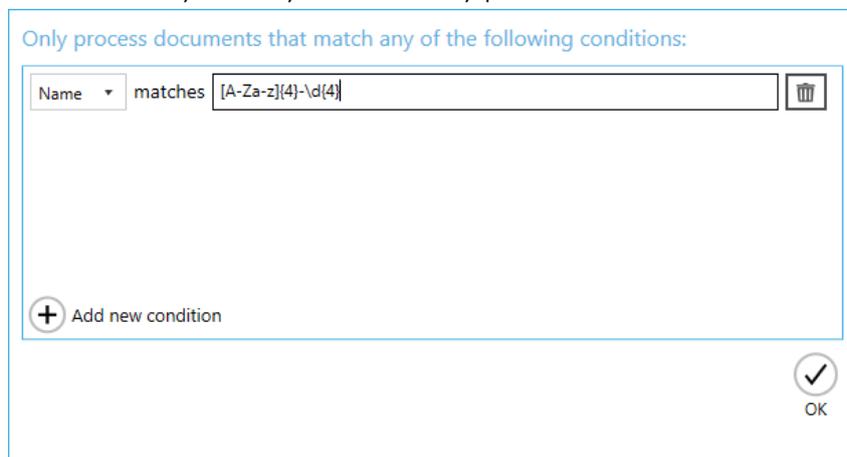
 OK

The location filters can be added through **Library > Library Settings > Filter Locations by Regular Expression**.

- Document Filtering

This can be useful if you want to only process documents with a certain naming convention.

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



Only process documents that match any of the following conditions:



 Add new condition

 OK

The document filters can be added through **Library > Document Settings > Filter Documents by Regular Expression**.

1.1.2 Pause and Stop during enumeration stage

The Pause and Stop (Abort) functionality can now be used during the document enumeration stage. Previously, it was only available during the Audit and the OCR stages.

Note, however, that the more cores you are using (**Library > Library Settings > Cores**), the longer it will take to Pause or Stop the enumeration.

1.2 Bug Fixes

1.2.1 Remove visible text

Remove visible text was not working for the Aquaforest OCR engine.

1.2.2 "Invalid URL" error when adding specific URLs to "Exclude Locations"

The issue occurred when adding URLs that contain spaces. This fix also addresses another issue where child URLs were not getting excluded if a root URL was added to the excluded locations.

1.2.3 Searchlight erroring out when processing .MSG files

This error occurred when a library that had .MSG files with no PDF attachments was processed. If after processing (auditing), a document was deleted from the SharePoint library and the library was run again, it would error out.

1.2.4 "Invalid URL" error when adding sites/site collections that have periods (.)

Searchlight could not add sites/sites collections that had periods (.) in them.
e.g. "https://test.sharepoint.com/sites/site.with.period"

1.2.5 Scheduler issues

When searchlight was set to run continuously for short intervals (e.g. every 5 mins), it stopped working after 1 or 2 days even though the service was still running.

1.2.6 Adding O365 Locations

When adding a new O365 site collection or site or document library, users were clicking on the "Find" button instead of the "Save" button.

The "Find" feature is to enumerate O365 site collections if a tenant admin URL is added. The admin URL is usually in the format: **https://{mysite}-admin.microsoft.com**

However, if a non-admin tenant URL is specified (i.e. normal site collection/site/document library URL) and "Find" was clicked, it was giving the impression of enumerating site collections without ever returning or giving an error message. This has now been fixed.

1.2.7 "Request uses too many resources" when processing very large lists

When processing lists with very large number of list items, Searchlight would fail at the enumeration stage with one of the following errors:

- The Request uses too many resources
- Too many requests

When searchlight retrieve items from the SharePoint, it did so in batches of 2,000. For SharePoint Document Libraries, this batch size works without any errors. However, it may not work for SharePoint Lists because each item in a List can have one or more attachments and as a result this batch size increases by the number of attachments (2000 * Average no. of attachments per list item). This causes the error above.

To fix this issue, you can increase the values of 'MaxResourcesPerRequest' and 'MaxObjectPaths' using PowerShell. Note, however, this only applies to SharePoint On-Premises.

To view the existing value for these settings, run the following command in PowerShell:

```
Get-SPWebApplication | %{$_.ClientCallableSettings}
```

To increase the values run the following commands:

```
$webApp = Get-SPWebApplication "<SITEURL>"
$webApp.ClientCallableSettings.MaxObjectPaths = 6000
$webApp.ClientCallableSettings.MaxResourcesPerRequest = 50
$webApp.Update()
```

A good value for 'MaxObjectPaths' is:

- ('listBatchSize' (see below) x Average no. of attachments per List Item) - if the error is generated when enumerating documents from a SharePoint List
- a value greater than 'libraryBatchSize' (see below) - if the error is generated when enumerating documents from a SharePoint Document Library

However, if you are using SharePoint Online (O365) or if the above solution is not feasible for you, there are now 2 new settings in the Searchlight.config file that can help with this issue:

- listBatchSize
- libraryBatchSize

The default value for both settings is 2000. Reducing the value of these settings will also fix the issue. You will have to reduce the value(s) by trial and error until the error goes away. Usually, a safe value is ('MaxObjectPaths' / Max no. attachments in the list items).

Note, however, the smaller the value for listBatchSize and libraryBatchSize, the longer the enumeration will take.

Make sure you restart the Searchlight service after making changes to Searchlight.config.

2 Version 1.20

2.1 Enhancements

2.1.1 Process PDF attachments inside MSG files

In this version, PDF attachments inside MSG files can be processed. The attachments are OCR'd and replaced in the MSG files.

2.1.2 Alerts and Reports

Aquaforest Searchlight now has the ability to generate scheduled CSV reports to show statistics about the status of a library as a whole as well as show statistics about particular job runs (such as jobs that were run within a particular date range) to find out how many documents were successfully OCR'd and how many failed.

Users can setup a report to run daily, weekly, monthly, etc. and automatically send an email with the report attached.

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

Configuration

Action
Email
Trigger
Finish

Trigger

When do you want the task to start?
At 23:00, on the first Monday, Wednesday, and Saturday of the month.

Start: 13/10/2016 23:00:00

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

Day(s): 4,8

The: First Monday, Wednesday, Saturday

Advanced Settings

On Job Success
 Yes

On Job Error
 Yes

Previous Next

Users can also manually generate CSV report of previous job runs. To do so, go to the "Run Details" tab, select a run history and click on "Export to CSV".

2.1.3 64-bit

As of version 1.20, Aquaforest Searchlight is a 64-bit application which means it can now process larger sets of documents as well as large documents concurrently without running out of memory (as long as your system has enough physical memory).

2.1.4 Check service status periodically

In previous versions, if the Searchlight service crashed (e.g. due to out of memory), the status of a running job was still set to as running on the Dashboard. This was misleading as users would not know the job had stopped unless they manually checked the status of the Searchlight service in the task manager.

In this release, a feature has been added 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 interval for checking the service is controlled by the "checkServiceEvery" option in Searchlight.config file. The default is to check the service every 60 minutes.

2.1.5 Ignore errors when enumerating folders

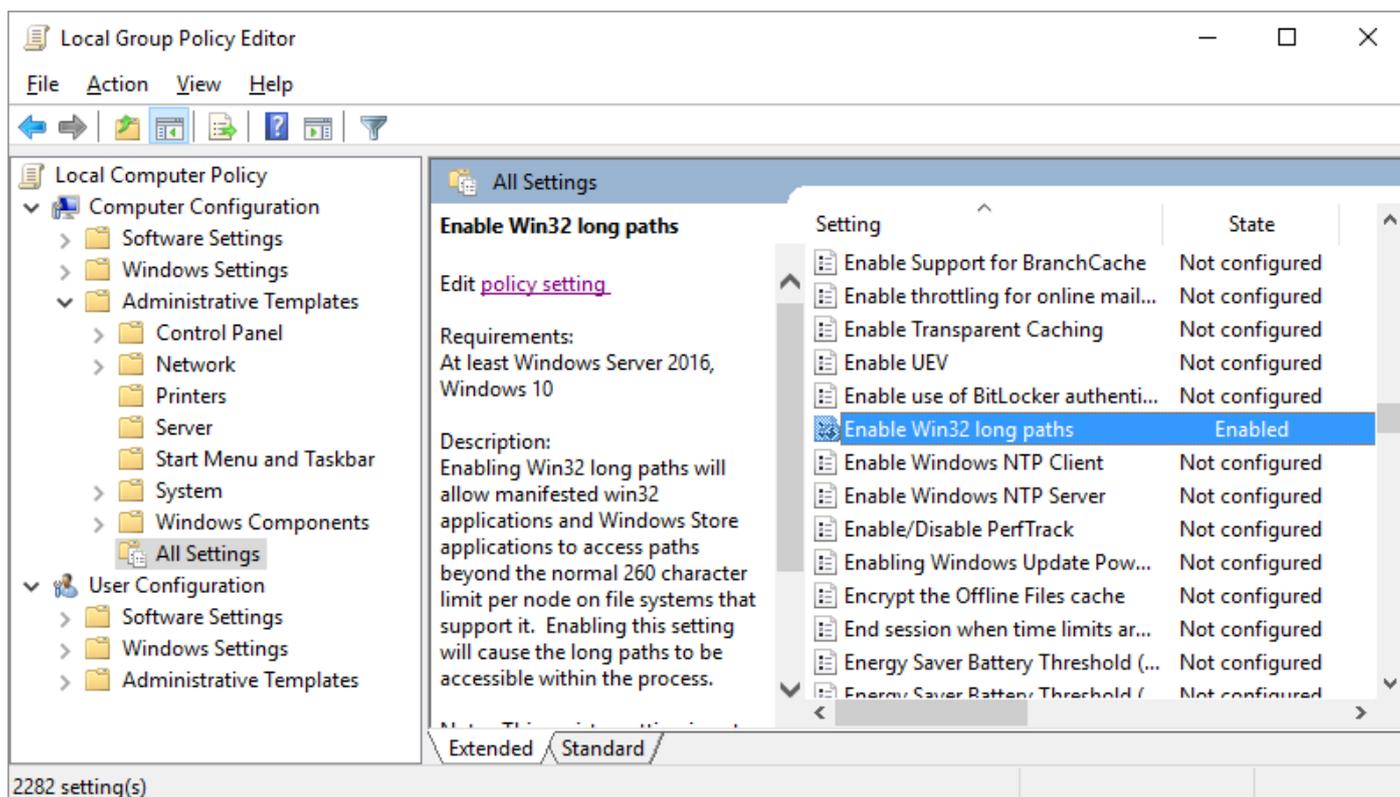
Folders that can't be enumerated due to permissions restrictions, long path errors, etc. can now be skipped instead of failing the whole job. This is controlled by the "skipEnumerationErrors" setting in the Searchlight.config file. This setting is only valid for File System sources.

2.1.6 Long Path support

When enumerating documents to process, Searchlight can come across documents that exceed the file path length enforced by windows. These files are skipped and not processed.

Starting from Windows 10 and Windows Server 2016, there is now support for long paths. However, long paths support is not enabled by default. You need enable the following policy to take advantage of this new feature.

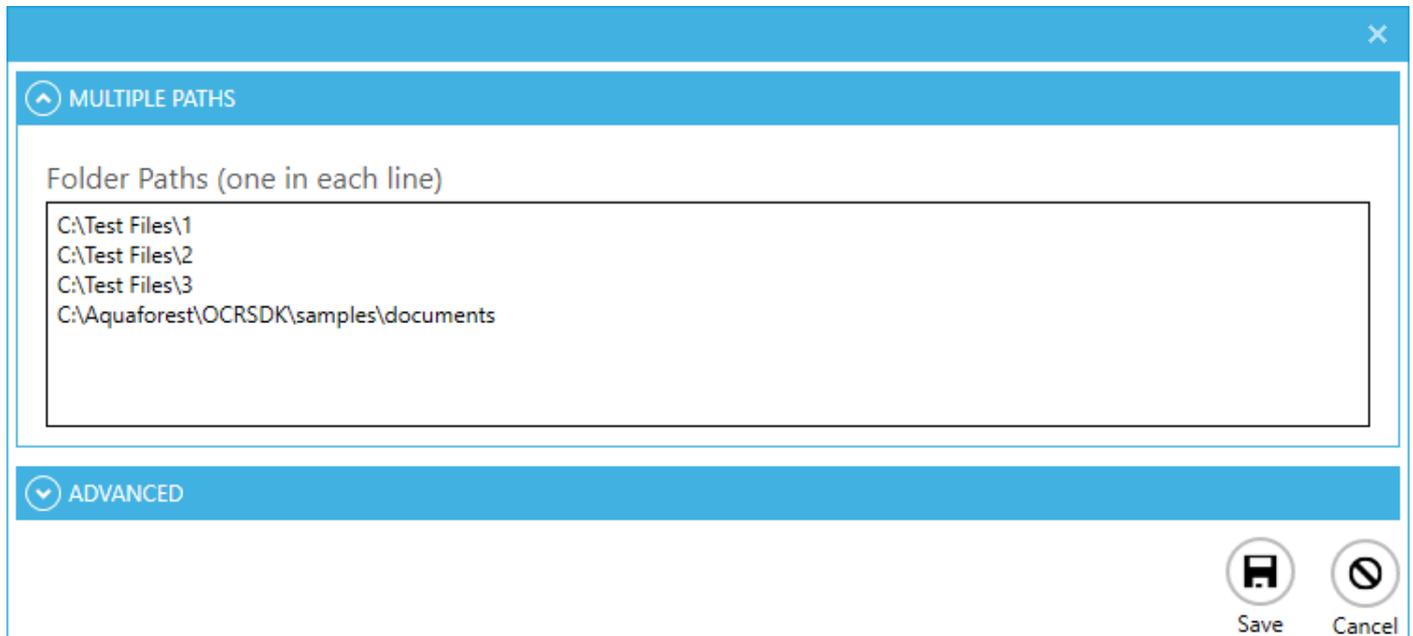
Open Global Policy Editor (**Start > Run > gpedit.msc**) and enable "**Enable Win32 long paths**".



Restart the Searchlight service after making changes to this policy.

2.1.7 Add multiple File System paths

Multiple file system paths can now be added all at once through the “Multiple Paths” expander as shown below:



2.1.8 Process PDF files with vector objects in native mode

PDF documents that contain only vector images (e.g. CAD drawings) can now be OCR'd natively. In previous versions, the PDF needed to be re-imaged before OCRing.

By default, pages that contain only vector objects are rasterized. Pages that do not have any images but contain vector objects as well as electronic text are skipped from rasterization. However, sometimes there can be a page that contains vector objects (CAD drawings) but its title may be in electronic text. To force rasterizing pages like these, there is a property called “PdfToImageForceVectorCheck” in the Properties.xml file of the OCR engine being used, which needs to be set to true. Note, however, that this is a global setting and will affect all document libraries using that particular OCR engine.

2.1.9 Font sizing

The sizing of OCR'd text added to PDF documents in native mode (i.e. without re-imaging) in the Extended OCR engine (IRIS) has been improved.

2.1.10 FIPS Compliancy

Aquaforest Searchlight as well as the OCR engines should now be FIPS compliant.

2.1.11 Temp Folder

The “Temp Folder Location” specified in the “Document Settings” tab will also be used to temporarily store OCR'd documents in addition to downloaded documents.

2.1.12 Force error when page exceeds pixel limit

A new setting has been added to 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). This is controlled by the "failOnPixelLimit" setting in the Searchlight.config file. 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

2.1.13 Retries

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.

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

2.1.14 Parallel Enumeration

When enumerating documents from large SharePoint libraries, Aquaforest Searchlight partitions the retrieval so that the documents are retrieved in chunks. In this release, these chunks can be retrieved in parallel which can significantly speed up enumeration. The maximum number of chunks that can be retrieved at once is controlled by the "enumerationMaxParallelism" setting in the "Searchlight.config" file. Note, however, that the maximum value will be limited to the maximum cores your license permits.

2.1.15 Audit page limit

A new feature has been added to limit the number of pages per document to audit. This can be beneficial for documents with lots of pages as it will speed up the audit process. This feature is controlled by the "maxAuditPageCount" in the Searchlight.config file. The default value for this setting is 0, which means that Searchlight will audit all pages of each document.

2.1.16 Check-in comment for failed documents

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 fails to OCR, no comment is added.

To force Searchlight to add a comment to the original non-OCRed document in SharePoint, specify a comment in the "sharePointFailCheckinComment" setting in the Searchlight.config file.

2.2 Bug Fixes

2.2.1 Scheduler

The scheduler option “Continuous every x days” did not work properly. This has now been fixed.

2.2.2 UI crash

Fixed issue where the UI crashed if values from drop-down menus were selected when there were no document library in Aquaforest Searchlight.

3 Version 1.10

3.1 Enhancements

3.1.1 Updated Extended OCR engine

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

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

3.1.2 Re-image PDF

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

3.1.3 Convert PDF to PDF/A

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

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

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

Dashboard Library Settings Help & Support

Status Library Settings Document Settings Archive Settings OCR Settings Run Details

PDF Selection

Process PDF Documents
Yes

Image Only PDFs
Yes

Partially Searchable
Yes

Fully Searchable
No

Hidden Text
Yes

TIFF Selection

Process TIFF Files
No

Delete Original TIFF
No

BMP Selection

Process BMP Files
No

Delete Original BMP
No

JPEG Selection

Process JPEG Files
No

Delete Original JPEG
No

PNG Selection

Process PNG Files
No

Delete Original PNG
No

Temp Folder Location:
C:\Aquaforest\Searchlight\temp

Filter Settings

Filter Rule:
No Filter

From: 12/02/2015 To: 12/02/2015

Exclude Specific Documents

Document Error Settings

Document Error Rule:
Take no Action

Document Error Location:

3.1.5 Exclude specific documents

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

Dashboard Library Settings Help & Support

Status Library Settings Document Settings Archive Settings OCR Settings Run Details

PDF Selection

Process PDF Documents
Yes

Image Only PDFs
Yes

Partially Searchable
Yes

Fully Searchable
No

Hidden Text
Yes

TIFF Selection

Process TIFF Files
No

Delete Original TIFF
No

BMP Selection

Process BMP Files
No

Delete Original BMP
No

JPEG Selection

Process JPEG Files
No

Delete Original JPEG
No

PNG Selection

Process PNG Files
No

Delete Original PNG
No

Temp Folder Location:
C:\Aquaforest\Searchlight\temp

Filter Settings

Filter Rule:
No Filter

From: 12/02/2015 To: 12/02/2015

Exclude Specific Documents

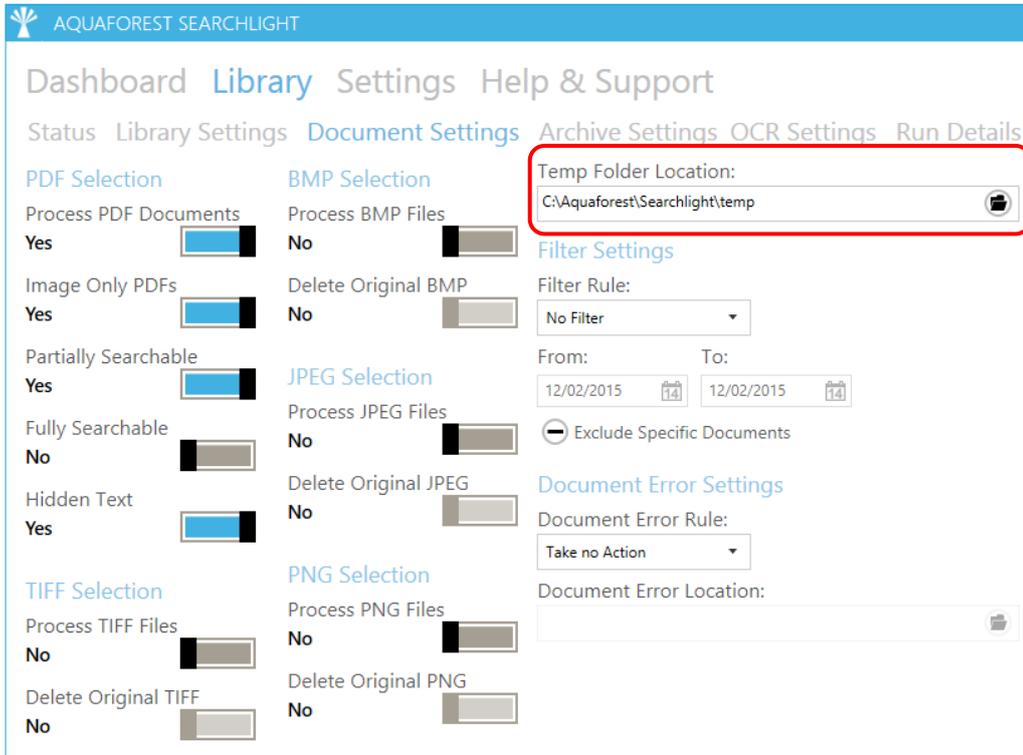
Document Error Settings

Document Error Rule:
Take no Action

Document Error Location:

3.1.6 Temp Location

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

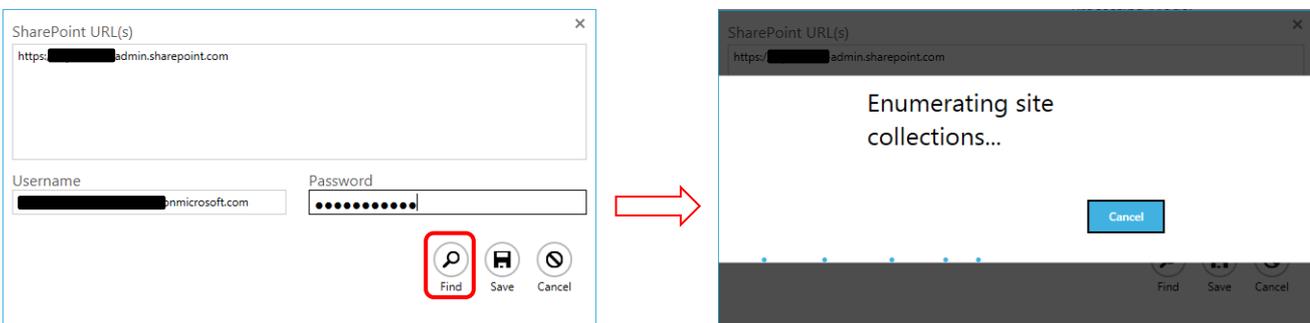


3.1.7 Active Directory Federation Service (AD FS) login

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

3.1.8 Enumerate site collections

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



3.1.9 Retrieve documents from SharePoint lists that exceed the List View Threshold

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

3.1.10 Audit and OCR documents one by one

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

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

3.1.11 Default OCR settings

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

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

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

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

3.1.12 Ignore previously OCRed documents

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

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

3.1.13 Skip checked-out documents

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

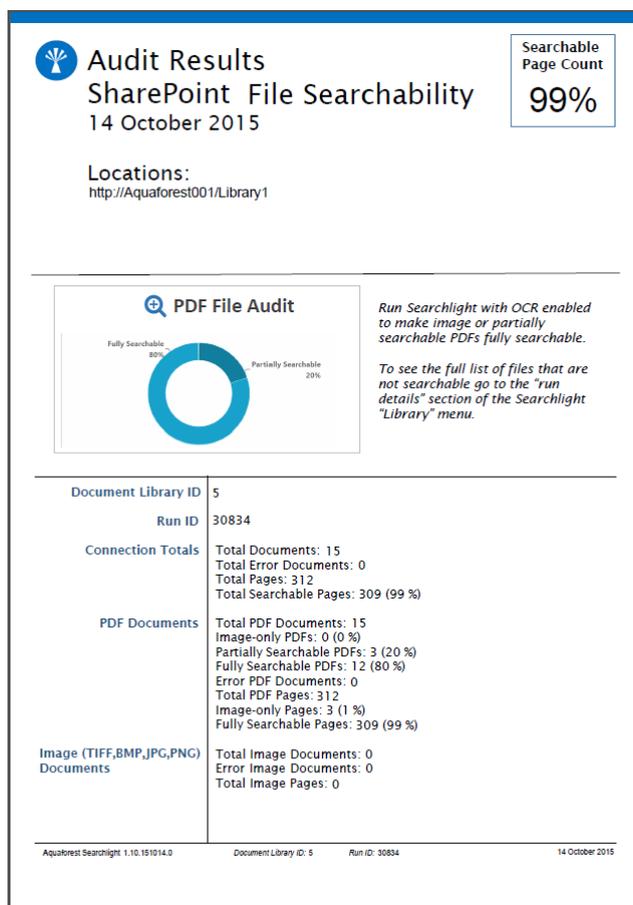
3.1.14 Retain Approval Status

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

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

3.1.15 Audit Chart

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



3.1.16 Performance

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

3.1.17 Database Locks

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

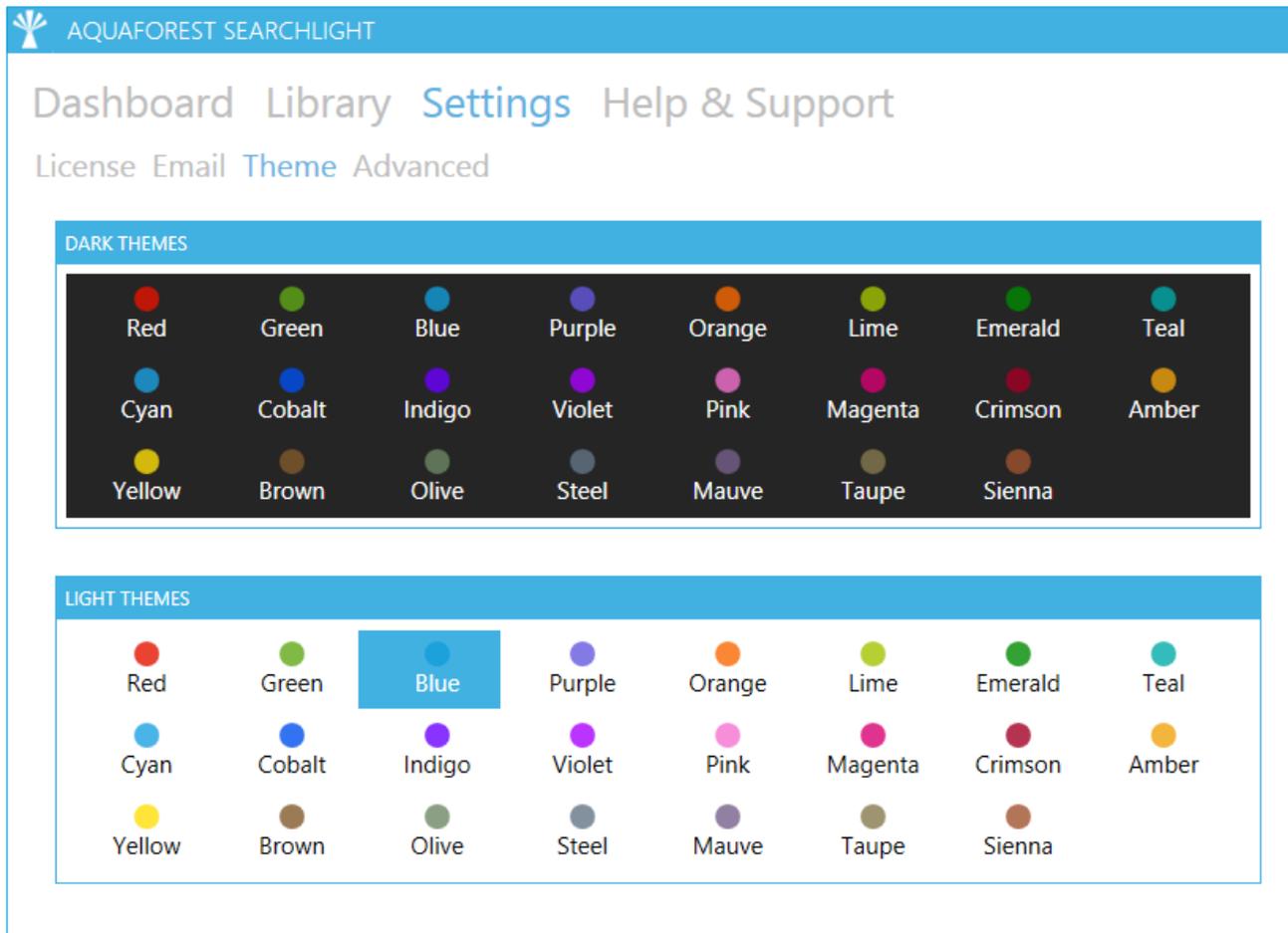
3.1.18 UI Changes

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

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

3.1.19 New Themes

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



4 Version 1.05

4.1 Enhancements

4.1.1 Add Multiple SharePoint URLs

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

SharePoint URL(s)

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

Username: username

Password:

Save Cancel

4.1.2 Download Progress

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

4.1.3 Download Retries

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

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

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

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

4.1.4 Database Update Retries

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

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

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

This config setting can be found in the "Searchlight.config" file located at:

"[installation path]\config\Searchlight.config".

4.1.5 Form-based authentication

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

4.1.6 Remove Hidden Text

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

4.1.7 Remove Visible Text

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

To enable this feature:

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

4.1.8 Retain Creation/Modified Date/User

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

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

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

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

4.1.9 Multicore support

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

4.2 Bug Fixes

4.2.1 SharePoint Template Types

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