
Aquaforest Searchlight Reference Guide



Version 1.10
February 2016

Contents

| | | |
|-------|---|---|
| 1 | Product Overview | 1 |
| 1.1 | The Business Problem: Documents that are not searchable. | 1 |
| 1.2 | The Solution: Aquaforest Searchlight | 1 |
| 1.3 | Supported Document Stores..... | 2 |
| 1.4 | Supported Operating Systems | 2 |
| 1.5 | Aquaforest Searchlight Modules | 2 |
| 1.5.1 | Multi-Core Module (Included as Standard)..... | 2 |
| 1.5.2 | Extended OCR Module (Included as Standard)..... | 2 |
| 1.5.3 | Extended OCR Asian Language Module (Extra Cost) | 2 |
| 1.5.4 | Extended OCR Advanced Compression Module (Extra Cost) | 2 |
| 2 | Installation and Licensing | 3 |
| 2.1 | Installation | 3 |
| 2.2 | System Requirements | 3 |
| 2.3 | SharePoint Online (Office 365) System Requirements | 3 |
| 2.4 | Microsoft SharePoint Requirements..... | 3 |
| 2.5 | Licensing..... | 3 |
| 2.5.1 | Entering License Keys | 3 |
| 2.6 | Trial License Mode | 4 |
| 3 | Searchlight Architecture | 4 |
| 3.1 | Processing Model..... | 4 |
| 3.1.1 | Audit and Candidate Identification | 4 |
| 3.1.2 | Searchability Status | 4 |
| 3.1.3 | Supported Formats | 4 |
| 3.2 | Aquaforest Searchlight Windows Service..... | 4 |
| 4 | Document Library Management | 5 |
| 4.1 | Creating a Document Library | 5 |
| 4.1.1 | Library Settings | 5 |
| 4.1.2 | SharePoint URL format | 6 |
| 4.1.3 | Document Settings | 6 |
| 4.1.4 | Document Archive Settings | 7 |
| 4.1.5 | OCR Settings..... | 7 |
| 4.1.6 | Scheduler..... | 8 |
| 4.2 | Editing a Document Library | 9 |

| | | |
|--------|---|----|
| 4.3 | Deleting a Document Library | 9 |
| 4.4 | Processing/Running a Document Library | 9 |
| 4.5 | Review Audits & Conversions | 9 |
| 4.6 | Reports..... | 9 |
| 4.6.1 | Error documents..... | 9 |
| 4.6.2 | Library Status | 9 |
| 4.6.3 | Generating Log files | 10 |
| 5 | Microsoft SharePoint Document Stores..... | 12 |
| 5.1 | Versioning..... | 12 |
| 5.2 | Archiving | 12 |
| 5.3 | Metadata..... | 12 |
| 5.4 | Security | 12 |
| 6 | File System Document Stores..... | 13 |
| 6.1 | File Name Length | 13 |
| 6.2 | File Access Permissions | 13 |
| 6.3 | Archiving | 13 |
| 6.4 | Metadata..... | 13 |
| 7 | Aquaforest Searchlight Modules..... | 14 |
| 7.1 | Multi-Core Module (Included with the standard product) | 14 |
| 7.2 | Extended (IRIS) OCR Module (Included with the standard product) | 14 |
| 7.3 | Extended OCR Advanced Compression (Extra Cost Option) | 14 |
| 8 | Aquaforest Searchlight User Interface | 15 |
| 8.1 | Welcome Screen | 15 |
| 8.2 | Aquaforest Searchlight Dashboard | 16 |
| 8.2.1 | Dashboard Fields | 16 |
| 8.3 | Library | 17 |
| 8.3.1 | Library Status | 17 |
| 8.3.2 | Library Settings | 18 |
| 8.3.3 | Document Settings | 19 |
| 8.3.4 | Document Archive Settings | 21 |
| 8.3.5 | Aquaforest OCR Settings | 22 |
| 8.3.6 | Extended OCR Settings..... | 25 |
| 8.3.7 | Run Details | 29 |
| 8.3.8 | Run Details Context Menu..... | 29 |
| 8.3.9 | Scheduler Settings | 30 |
| 8.3.10 | Alert Settings..... | 31 |
| 8.4 | Support..... | 31 |

| | | |
|-------|--|----|
| 8.5 | Diagnostic Tool..... | 31 |
| 8.6 | Settings | 33 |
| 8.6.1 | License Settings..... | 33 |
| 8.6.2 | Email Settings..... | 33 |
| 8.6.3 | Themes | 34 |
| 8.6.4 | Advanced Settings..... | 35 |
| 9 | Aquaforest Searchlight Installation Files..... | 36 |
| 10 | Acknowledgements..... | 37 |

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, Windows File Systems and File System-based Document Management Systems such as Worldox.

Aquaforest Searchlight revolves around **Document Libraries**. A Document Library in Aquaforest Searchlight is an object that has references to one or more Document Management System locations. Aquaforest Searchlight then monitors these locations and makes sure that the TIFF, BMP, JPEG, PNG and PDF documents contained in them are made searchable.

1.1 The Business Problem: Documents that are not searchable.

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

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

- Scanned TIFF Files
- Image PDF Files
- Faxes

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

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

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

1.2 The Solution: Aquaforest Searchlight

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




1.3 Supported Document Stores

- Windows File Systems
- SharePoint 2010
- SharePoint 2013
- SharePoint Online (Office 365)
- Worldox and other Document Management Systems where files are held in a Windows file system

1.4 Supported Operating Systems

Windows 2008 Server, Windows 2012 Server, Windows 7, Windows 8

1.5 Aquaforest Searchlight Modules

| | |
|--|--|
| <p>Aquaforest</p>  | <p>1.5.1 Multi-Core Module (Included as Standard) Take full advantage of multi-core architectures for maximum performance.</p> <p>1.5.2 Extended OCR Module (Included as Standard) Adds the IRIS OCR Engine for more comprehensive Language support (over 100 languages) and enhanced accuracy for certain document types.</p> <p>1.5.3 Extended OCR Asian Language Module (Extra Cost) Adds support for Korean, Japanese, Simplified Chinese & Traditional Chinese.</p> <p>1.5.4 Extended OCR Advanced Compression Module (Extra Cost) Adds support for advanced Mixed Raster Compression that can dramatically reduce the size of color scanned documents.</p> |
|--|--|

2 Installation and Licensing

2.1 Installation

The product is installed via the aquaforest_searchlight.msi install package that can be downloaded from Aquaforest's download page: http://aquaforest.com/en/searchable_pdf_download.asp

2.2 System Requirements

| | |
|------------------------------------|--|
| Supported Operating Systems | Windows 7 Windows 8 Windows 10 Windows Server 2008 R2 Windows Server 2012 |
| Disk Space | 480 MB |
| Memory | Minimum 4GB (recommended 8GB) |
| Visual C++ Redistributable | Visual C++ 2010 Redistributable (x86 x64) and Visual C++ 2012 Redistributable (x86) |
| .NET Framework | 3.5 and 4.5 |

2.3 SharePoint Online (Office 365) System Requirements

| | |
|------------------------------------|---|
| Supported Operating Systems | Windows 7 SP1 and above Windows 8 Windows 10 Windows Server 2008 R2 SP1 and above Windows Server 2012 |
| Additional tools | SharePoint Server Client Components SDK |

2.4 Microsoft SharePoint Requirements

In order to make use of the SharePoint functionality, Microsoft SharePoint 2010 or 2013 (Standard or Enterprise) is required. Note that the metadata service is required before the Entity Extraction module can be used.

2.5 Licensing

2.5.1 Entering License Keys

License keys may be entered by using the "settings" tab from the Aquaforest Searchlight application. Without a purchased license key the product operates in trial mode. The trial license provides access to the following modules:

- Aquaforest OCR
- Extended OCR enabling users to choose the IRIS OCR engine.
- Asian Language Support.
- Multi-Core Module for parallel processing.

Further Modules are also available upon request, these are:

- Intelligent High Quality Compression
- Asian Languages OCR support

2.6 Trial License Mode

In this mode there is a limit of 100 documents that may be converted although there is no limit on the number of documents that may be audited.

3 Searchlight Architecture

3.1 Processing Model

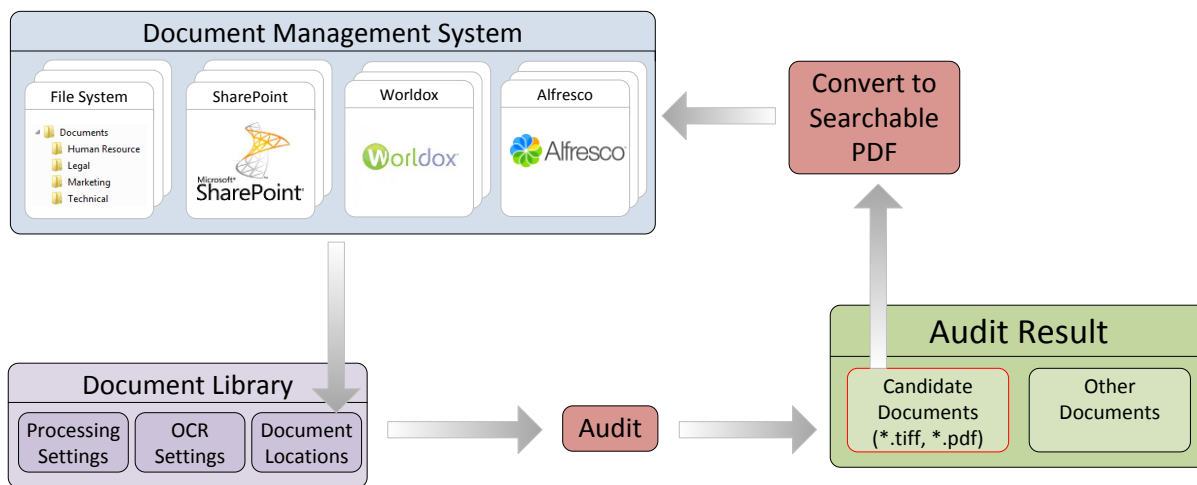


FIGURE 3.1: PROCESSING MODEL

3.1.1 Audit and Candidate Identification

Before processing a document library, Aquaforest Searchlight will perform an Audit (analysis) on the document library in order to determine which documents are candidates for processing by examining each document's searchability status (see next section) and the document library's processing settings.

3.1.2 Searchability Status

This determines whether a PDF file needs to be processed according to the Conversion Rules. Status values include Unknown, Searchable, Error, Partially Searchable amongst others.

3.1.3 Supported Formats

Aquaforest Searchlight currently supports only TIFF, BMP, JPG, PNG and PDF documents as input. As a result, candidate documents will always be of TIFF, BMP, JPG, PNG and/or PDF format.

3.2 Aquaforest Searchlight Windows Service

The service controls the execution of all job runs in Aquaforest Searchlight. It is used by the scheduler and enables the monitoring and processing of document libraries at regular time intervals without interfering with other work being performed on the machine it is installed in.

4 Document Library Management

Aquaforest Searchlight revolves around the concepts of document libraries, a document library is a set of documents which Aquaforest Searchlight monitors and ensures are converted where required in accordance with the Document Library Conversion Rules. This section will discuss the different concepts associated with Document Libraries.

4.1 Creating a Document Library

Creating Document Libraries in Aquaforest Searchlight is managed by a wizard. This wizard can be launched by clicking the "Create New Library" button on the Dashboard tab. The wizard contains five pages that will guide you through the set up.

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 sections on Library Settings and OCR Settings for detailed description of each of the settings.

4.1.1 Library Settings

The screenshot shows the 'Library Settings' page of the Aquaforest Searchlight wizard. The interface has a blue header with the application name and navigation tabs: Dashboard, Library (selected), Settings, and Help & Support. A dropdown menu in the top right shows 'File System'. Below the tabs, there are links for Status, Library Settings (selected), Document Settings, Archive Settings, OCR Settings, Run Details, Scheduler, and Alerts. The main settings area includes: 'Library Name' with a text box containing 'File System'; 'Library Type' with a dropdown menu set to 'File System'; 'Locations' with a list box containing 'C:\Test Files' and an 'Add new Location' button; 'Choose Library Icon' with a blue folder icon; 'Processing Mode' with radio buttons for 'Audit Only' and 'Audit and OCR' (selected); 'Cores' with a dropdown menu set to '1'; 'SharePoint Settings' section with 'If Versioning is Off' set to 'Turn versioning on', 'Publish Major Version' set to 'Yes' with a toggle switch, and a 'Check-In Comment' text box containing 'OCR'ed by Aquaforest Searchlight'; and an 'Exclude Locations' button. At the bottom right are 'Delete', 'Refresh', and 'Save' buttons. The bottom left shows 'Service Status : Running'.

This section enables users to specify the locations or URL of the documents they wish to monitor. To add a collection of documents to the document library follow the steps below:

- Select the Document Management system type you wish to add, this will be SharePoint, Office 365 or the Windows File System.
- When that is selected, a group box will be displayed that will allow the user to provide the location of the document collection and the security credentials (Username and Password).

- This will then display all the document collections on folders found at that particular location allowing the user to select which documents to add to the document library.
- If the documents are selected, the user should click on the "Add Another Location" button to add more documents to the library.
- If the User is happy with the documents added, a next button is provided to move to the next stage of the library set up.

4.1.2 SharePoint URL format

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

SharePoint 2010, actual URL:

[http\(s\)://SharePoint2010/site/myLibrary/myForms/AllItems.aspx](http(s)://SharePoint2010/site/myLibrary/myForms/AllItems.aspx)

Supported URL format in Searchlight:

[http\(s\)://SharePoint2010/site/myLibrary](http(s)://SharePoint2010/site/myLibrary)

SharePoint 2013, actual URL:

[http\(s\)://SharePoint2013/site/Library/_layouts/15/start.aspx#/mylibrary/Forms/AllItems.aspx](http(s)://SharePoint2013/site/Library/_layouts/15/start.aspx#/mylibrary/Forms/AllItems.aspx)

Supported URL format in Searchlight:

[http\(s\)://SharePoint2013/site/Library/mylibrary](http(s)://SharePoint2013/site/Library/mylibrary)

4.1.3 Document Settings

This tab enables the user to specify more rules and criteria for the selection of documents to be added to the Document Library. These criteria include date created, date modified, file types to add, and document count limit.

The screenshot displays the 'AQUAFORREST SEARCHLIGHT' application window with the 'Document Settings' tab selected. The interface is organized into several sections:

- Navigation:** Dashboard, Library, Settings (selected), Help & Support. Sub-tabs include Status, Library Settings, Document Settings (selected), Archive Settings, OCR Settings, Run Details, Scheduler, and Alerts.
- PDF Selection:** Process PDF Documents (Yes), Image Only PDFs (Yes), Partially Searchable (Yes), Fully Searchable (Yes), Hidden Text (Yes).
- TIFF Selection:** Process TIFF Files (Yes), 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).
- Filter Settings:** Temp Folder Location (C:\Aquaforest\Searchlight\temp), Filter Rule (No Filter), From/To dates (12/06/2014), and an option to Exclude Specific Documents.
- Document Error Settings:** Document Error Rule (Take no Action), Document Error Location.
- Advanced Settings:** Retry (Yes), OCR Document Limit (55), Retain Creation Date (No), Retain Modified Date (No), Retain Created By (No), and Retain Modified By (No).
- Service Status:** Running.
- Buttons:** Refresh and Save.

4.1.4 Document Archive Settings

This tab enables archiving source files so as to have a backup before OCR is applied to them.

The screenshot shows the 'Document Archive Settings' tab in the Aquaforest Searchlight application. The interface includes a top navigation bar with 'Dashboard', 'Library', 'Settings', and 'Help & Support'. Below this is a sub-navigation bar with 'Status', 'Library Settings', 'Document Settings', 'Archive Settings' (highlighted), 'OCR Settings', 'Run Details', 'Scheduler', and 'Alerts'. The main content area is titled 'Document Archive Settings' and contains the following settings:

- Archive Rule:** A dropdown menu set to 'Copy to Archive Folder'.
- Archive Template:** A text input field containing '%FILENAME%%TIMESTAMP%.%EXT%'.
- Archive Location:** A text input field containing 'C:\TestFiles\Archive' with a folder icon button to its right.
- Archive source TIFFs to Archive Folder:** A toggle switch set to 'Yes'.
- Archive source PDFs to Archive Folder (even when versioning is on):** A toggle switch set to 'Yes'.

4.1.5 OCR Settings

These tabs will enable the user set the OCR Engine to their taste. Aquaforest Searchlight comes bundled with two OCR Engines and each of these engines has different settings.

4.1.5.1 Aquaforest OCR Engine Settings

This page allows the user to choose the settings suitable for the Aquaforest OCR engine.

The screenshot shows the 'Aquaforest OCR Engine Settings' tab in the Aquaforest Searchlight application. The interface includes a top navigation bar with 'Dashboard', 'Library', 'Settings', and 'Help & Support'. Below this is a sub-navigation bar with 'Status', 'Library Settings', 'Document Settings', 'Archive Settings', 'OCR Settings' (highlighted), 'Run Details', 'Scheduler', and 'Alerts'. The main content area is titled 'Aquaforest OCR Engine Settings' and contains the following settings:

- OCR Engine:** Two radio buttons, 'Aquaforest' (selected) and 'Extended (IRIS)'.
- General Settings:** A sub-section containing:
 - Auto Rotate:** A toggle switch set to 'Off'.
 - Despeckle:** A dropdown menu set to 'No Despeckle'.
 - Deskew:** A toggle switch set to 'Off'.
 - OCR Language:** A dropdown menu set to 'English'.
 - Remove Lines:** A toggle switch set to 'Off'.
 - Box Graphics:** A dropdown menu set to 'Remove Box Lines in OCR Processing'.
 - JBIG2 Compression:** A toggle switch set to 'Off'.
 - Stamps:** An empty text input field.
 - Advance Flags:** An empty text input field.

4.1.5.2 Extended OCR Engine Settings.

This page allows the user to choose the settings suitable for the extended OCR engine.

The screenshot shows the 'OCR Settings' page in the Aquaforest Searchlight application. The top navigation bar includes 'Dashboard', 'Library', 'Settings', and 'Help & Support'. Below this, a sub-navigation bar lists 'Status', 'Library Settings', 'Document Settings', 'Archive Settings', 'OCR Settings' (which is highlighted), 'Run Details', 'Scheduler', and 'Alerts'. The main content area is titled 'OCR Engine:' and shows two radio buttons: 'Aquaforest' and 'Extended (IRIS)', with 'Extended (IRIS)' selected. Below this, there are four tabs: 'General Settings' (selected), 'PDF Source Settings', 'TIFF Source Settings', and 'Advanced Preprocessing Settings'. The 'General Settings' section contains several controls: 'Auto Rotate' (set to 'Off'), 'Deskew' (set to 'No'), 'Remove Dark Borders' (set to 'No'), 'Keep Original Image' (set to 'No'), 'Despeckle' (set to 'No Despeckle'), 'Advanced Despeckle' (set to 'No Despeckle'), 'Remove White Pixels' (set to 'No'), 'Work Depth' (set to '1'), 'Interpolate' (set to 'No'), 'Interpolation Mode' (set to 'Normal'), and 'Interpolation Value' (set to '1'). On the right side, there is a 'Select Language(s)' dropdown menu with a list of languages: Czech, Danish, Dutch, English (checked), Esperanto, Estonian, Faroese, Fijian, Finnish, and French. The 'English' option is highlighted in blue.

4.1.6 Scheduler

The scheduler allows Aquaforest Searchlight to automate the monitoring and running of document libraries.

The screenshot shows the 'Scheduler' page in the Aquaforest Searchlight application. The top navigation bar is the same as in the previous screenshot. The sub-navigation bar lists 'Status', 'Library Settings', 'Document Settings', 'Archive Settings', 'OCR Settings', 'Run Details', 'Scheduler' (which is highlighted), and 'Alerts'. The main content area is titled 'Scheduler' and shows three radio buttons: 'Manual' (selected), 'Once per day', and 'Continuous'. Below these, there are three sections for scheduling: 1. 'Manual' section: A single radio button. 2. 'Once per day' section: A radio button with a time picker set to '15 : 31'. 3. 'Continuous' section: A radio button with a frequency picker set to '1' and a unit dropdown set to 'Hours'. Below this, there are two time pickers: 'Between' (set to '00 : 01') and 'And' (set to '23 : 59'). 4. 'Run once' section: A radio button with a date picker set to '12/02/2015' and a time picker set to '15 : 31'.

4.2 Editing a Document Library

Aquaforest Searchlight allows the user to edit the library locations, library settings and the OCR Settings at any time. These settings are accessible through the Library Tab.

4.3 Deleting a Document Library

Aquaforest Searchlight will allow a user to delete a document library at any time as long as it is not in a running state.

4.4 Processing/Running a Document Library

Aquaforest Searchlight processes a document library by first carrying out an audit to find out how many files need to be converted. After a successful audit, Aquaforest Searchlight can then start conversion. Starting a run can either be done by the user or the scheduler. If the Ad-Hoc option is selected during the scheduler set up, the user will have to manually click on the "Run" button on the dashboard.

4.5 Review Audits & Conversions

Since Aquaforest Searchlight processes document in an in place fashion, it is best to always know the exact state of the Document Library before conversion begins. This process also helps the user compare the state of the Document Library before and after conversion.

The audit details are made available to user via the Run Details tab under the Library Tab. The audit details provide information such as: the list of all the files in the library, the number of searchable files in the library, the library size and much more.

A user can also choose the number of past Audits that should still be kept in Aquaforest Searchlight, this option is provided in the Library Settings tab both in the Create New Library wizard and under the Library Drill-Down tab.

4.6 Reports

Aquaforest Searchlight provides reports that show statistics for particular document libraries, conversions as well as error and support reports. Reports can be generated in PDF or Excel formats. The following reports are included with Aquaforest Searchlight.


- Aquaforest Searchlight Summary Report – Shows a summary of all the document libraries and settings (email settings, license, etc.) currently configured as well as the support status
- Document Library Report – Shows all the statistics related to a particular document library including the number conversions performed, success rate, documents processed, etc.
- Audit Report – Shows the results of an audit
- Conversion Report – Shows the statistics of a particular conversion
- Conversion Error Report – Shows all the errors and exceptions that occurred during a particular conversion

4.6.1 Error documents

With Aquaforest Searchlight, error documents are easy to locate. The Run Details tab provides filtering options which enables users to only display files using set criteria.

4.6.2 Library Status

This report can be accessed via the Library Status tab.


AQUAFOREST SEARCHLIGHT

Dashboard
Library
Settings
Help & Support

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

STATISTICS

PDF Documents


Total PDF Documents: 15
Image-only PDFs: 0 (0 %)
Partially Searchable PDFs: 3 (20 %)
Fully Searchable PDFs: 12 (80 %)
Error PDF Documents: 0
Total PDF Pages: 312
Image-only Pages: 3 (1 %)
Fully Searchable Pages: 309 (99 %)

Image (TIFF,BMP,JPG,PNG) Documents

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

Library Totals

Total Documents: 15
Total Error Documents: 0
Total Pages: 312
Total Searchable Pages: 309 (99 %)

Report


LOG OUTPUT

Number of Image Pages: 1, Number of Searchable Pages 55.
Contains hidden text

Downloading 'http://aquaforest001/Library1/aaabbbccc.pdf'
Checking document: http://aquaforest001/Library1/aaabbbccc.pdf
Searchability Status : searchable, Number of Pages 56.
Contains hidden text

Document library statistics after audit:

PDF Documents
Total PDF Documents: 15
Image-only PDFs: 0 (0 %)
Partially Searchable PDFs: 3 (20 %)
Fully Searchable PDFs: 12 (80 %)
Error PDF Documents: 0
Total PDF Pages: 312
Image-only Pages: 3 (1 %)
Fully Searchable Pages: 309 (99 %)

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


Library Totals
Total Documents: 15
Total Error Documents: 0
Total Pages: 312
Total Searchable Pages: 309 (99 %)

14-Oct-2015 9:55:08: Audit ended

Service Status : Running

4.6.3 Generating Log files

Log files can be generated using the “Generate Log” button which can be found on the Run Details tab under the Library Tab. This enables users to capture the information generated by any given Run ID into a PDF file which can be sent to the support team. This provides a file by file assessment of a document library before and after a run is executed. It shows the files that have been converted successfully and those that have failed.


AQUAFOREST SEARCHLIGHT

Dashboard
Library
Settings
Help & Support

qa.tiff

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

Run History

Showing last 5 runs

| RUN ID | RUN DATE | PROCESSING MODE | AUDIT RESULTS | | CONVERSION RESULTS | | |
|--------|---------------------|-----------------|---------------|---------------|--------------------|----------------------|-----------------|
| | | | Status | For Documents | Status | Successful Documents | Error Documents |
| 30982 | 16/09/2015 17:50:50 | Audit Only | Completed | | Not Yet Converted | 0 | 0 |

Filter Status

☒ Completed
 ☒ Error
 ☒ Aborted

ok

cancel

Run Details

☒ Audit
 ☐ Conversion

| DOCUMENT PATH | AUDIT STATUS | FILE TYPE | LAST MODIFIED | PAGES | CONVERSION ! |
|--|--------------|-----------|-----------------------|-------|--------------|
| C:\Test Files\qa\SearchlightQASets\qa\221112\split\chk_images001937.tif | imageonly | .TIF | 11/23/2012 4:28:22 PM | 1 | Audited |
| C:\Test Files\qa\SearchlightQASets\qa\221112\split\chk_images008999.tif | imageonly | .TIF | 11/23/2012 5:12:02 PM | 1 | Audited |
| C:\Test Files\qa\SearchlightQASets\qa\240113\ccube.tiff server vanilla bui | imageonly | .TIF | 7/13/2012 9:51:40 AM | 68 | Audited |
| C:\Test Files\qa\SearchlightQASets\qa\221112\split\chk_images008989.tif | imageonly | .TIF | 11/23/2012 5:11:56 PM | 1 | Audited |
| C:\Test Files\qa\SearchlightQASets\qa\221112\split\chk_images008952.tif | imageonly | .TIF | 11/23/2012 5:11:34 PM | 1 | Audited |
| C:\Test Files\qa\SearchlightQASets\qa\221112\split\chk_images007911.tif | imageonly | .TIF | 11/23/2012 5:01:53 PM | 1 | Audited |
| C:\Test Files\qa\SearchlightQASets\qa\221112\split\chk_images007959.tif | imageonly | .TIF | 11/23/2012 5:02:18 PM | 1 | Audited |
| C:\Test Files\qa\SearchlightQASets\qa\221112\split\chk_images007961.tif | imageonly | .TIF | 11/23/2012 5:02:19 PM | 1 | Audited |
| C:\Test Files\qa\SearchlightQASets\qa\221112\split\chk_images001541.tif | imageonly | .TIF | 11/23/2012 4:27:33 PM | 1 | Audited |

3

+

-

Limit 500

Generate Log

View Full Log

Reload

Service Status : Running

5 Microsoft SharePoint Document Stores

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

5.1 Versioning

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

5.2 Archiving

To avoid making inadvertent changes to the source document, Aquaforest Searchlight will ensure the archiving of source documents if versioning is turned off in SharePoint. Archiving is the process of copying over the source documents to an archive location specified by the user before performing any sort of processing on them.

5.3 Metadata

Metadata will be used to store the "Searchability" status of documents. Checks are made to determine if a document has been modified since it was last OCR'd and therefore may require re-processing.

5.4 Security

The user will have to provide the security credentials of the SharePoint library they wish to monitor during configuration.

6 File System Document Stores

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

6.1 File Name Length

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

6.2 File Access Permissions

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

6.3 Archiving

To avoid any loss of document, Aquaforest Searchlight is created to enforce archiving of input document. This is the process of copying over input documents to an archive location specified by the user before processing them.

6.4 Metadata

Metadata will be held in the Aquaforest Searchlight database to store details of the searchability status of a document. Checks are made to determine if a document has been modified since it was last OCR'd and therefore may require re-processing.

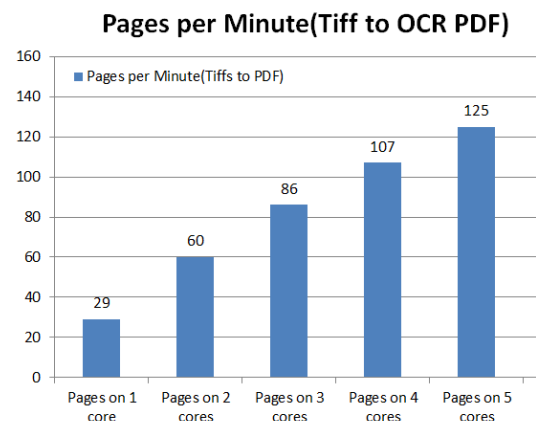
7 Aquaforest Searchlight Modules

7.1 Multi-Core Module (Included with the standard product)

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 10 files in parallel thereby reducing the time needed to execute a job.

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



7.2 Extended (IRIS) OCR Module (Included with the standard product)

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

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

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

- IRIS OCR Engine providing enhanced recognition
- New PDF Rasterizer component
- DOCX Output option
- Improved RTF Output
- CSV and Spreadsheet ML output options
- Supports over 100 Languages.
- Optional Asian Language Support
- Support for multiple languages within a single document from the same alphabet - e.g. French+ German + Italian

See section 8.3.5 for a detailed description of all the available Extended OCR settings.

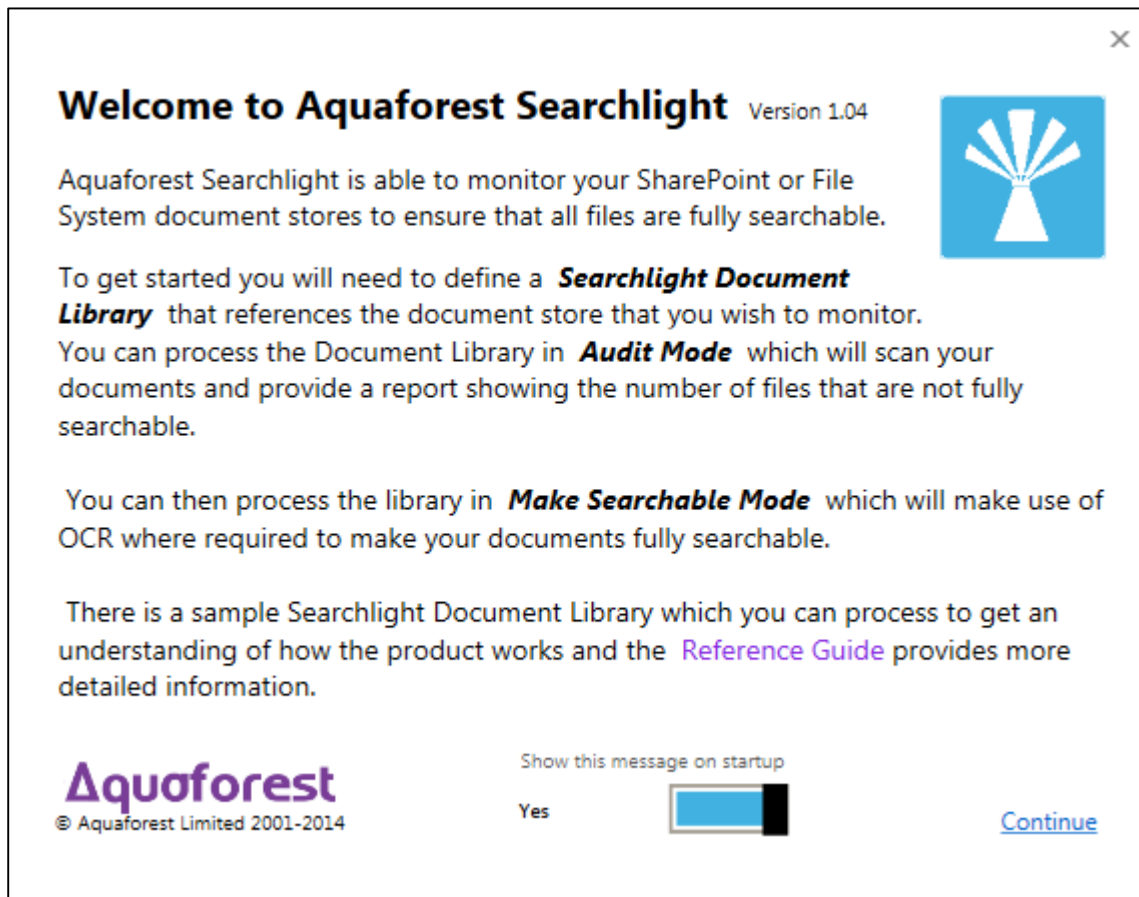
7.3 Extended OCR Advanced Compression (Extra Cost Option)

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. See section 8.3.6 for a detailed explanation of the settings available for Advanced Compression.

8 Aquaforest Searchlight User Interface

8.1 Welcome Screen

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



8.2 Aquaforest Searchlight Dashboard

Aquaforest SEARCHLIGHT




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

+ Create new library

| NAME | LIBRARY TYPE | LAST RUN | SCHEDULE | SEARCHABILITY | RUN STATUS |
|----------|--------------|----------------------|------------|---------------|------------|
| test one | File System | 02-Aug-2013 17:28:11 | Continuous | 97.8 % | Completed |

The dashboard gives a summary of the status of any document libraries currently processing or scheduled to process. A document library's previous conversions can be reviewed by selecting it from the list.


8.2.1 Dashboard Fields

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

8.3 Library

8.3.1 Library Status

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

 AQUAFOREST SEARCHLIGHT

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

Local

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

STATISTICS

PDF Documents

Total PDF Documents: 15

Image-only PDFs: 0 (0 %)

Partially Searchable PDFs: 3 (20 %)

Fully Searchable PDFs: 12 (80 %)

Error PDF Documents: 0

Total PDF Pages: 312

Image-only Pages: 3 (1 %)

Fully Searchable Pages: 309 (99 %)

Image (TIFF,BMP,JPG,PNG) Documents

Total Image Documents: 0

Error Image Documents: 0

Total Image Pages: 0


Library Totals

Total Documents: 15

Total Error Documents: 0

Total Pages: 312

Total Searchable Pages: 309 (99 %)

Report 

LOG OUTPUT

Number of Image Pages: 1, Number of Searchable Pages 55.
Contains hidden text

Downloading 'http://aquaforest001/Library1/aaabbbccc.pdf'
Checking document: http://aquaforest001/Library1/aaabbbccc.pdf
Searchability Status : searchable, Number of Pages 56.
Contains hidden text

Document library statistics after audit:

PDF Documents
Total PDF Documents: 15
Image-only PDFs: 0 (0 %)
Partially Searchable PDFs: 3 (20 %)
Fully Searchable PDFs: 12 (80 %)
Error PDF Documents: 0
Total PDF Pages: 312
Image-only Pages: 3 (1 %)
Fully Searchable Pages: 309 (99 %)

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


Library Totals
Total Documents: 15
Total Error Documents: 0
Total Pages: 312
Total Searchable Pages: 309 (99 %)

14-Oct-2015 9:55:08: Audit ended


Service Status : Running

Aquaforest Searchlight 1.10 Reference Guide

Page 17

 Aquaforest

8.3.2 Library Settings


AQUAFOREST SEARCHLIGHT

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

Site3_Library3

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


Please enter a Library Name:

Library Type:

Locations: + Add new Location

https://shrevin-hp/sites/sietcollectiontest/teamsite3/site3_library3

User Name:
Password:

Choose Library Icon:


Processing Mode:
☐ Audit Only
☒ Audit and OCR

Cores:

SharePoint Settings
If Versioning is Off:

Publish Major Version:
Yes ☒

Check-In Comment:

☒ Exclude Locations

Delete Refresh Save

Service Status : Stopped

| Setting | Description |
|------------------------------|--|
| Document Library Name | Name/Title/Description of the document library |
| Document Library Type | The type of the document library: File System SharePoint Office 365 |
| Locations | One or more locations to be processed. |
| Choose Library Icon | Choose an icon to associate to the library. |
| Processing Mode | <ul style="list-style-type: none"> Audit Only Analyse the document library to find out the documents that need to be converted without actually converting them Audit & OCR Perform audit on the document library and OCR the documents that have been identified as candidates for processing |
| Cores | This determines the maximum number of CPU cores that will be used when running the job. |
| SharePoint Versioning | This setting can be used to automatically turn versioning on. |
| Publish Major Version | Publish major version after OCR |
| SharePoint Exclude Locations | A set of URLs that will not be included in processing. |
| Check-in Comment | The check-in comment applied to the updated SharePoint file version. |

8.3.3 Document Settings

AQUAFOREST SEARCHLIGHT

Dashboard Library **Settings** Help & Support Demo

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

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:

Advanced Settings

Retry: **No**

OCR Document Limit: 0

Retain Creation Date: **No**

Retain Modified Date: **No**

Retain Created By: **No**

Retain Modified By: **No**

Service Status : Running

Refresh Save

| 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. |
| 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. |
| 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 |


| Setting | Description |
|----------------------------|---|
| 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 |
| Temp Folder Location | The folder used to save documents temporarily for Audit and OCR processing |
| Filter Rule | Select documents by modified date or creation date |
| Exclude Specific Documents | Exclude specific documents that will be ignored during processing |
| Document Error Rule | The operation to perform if a document fails to process: <ul style="list-style-type: none"> Copy to error folder Move to error folder (for file system library type only) |
| Document Error Location | The path of the error location |
| 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) |

8.3.3.1 Retain Creation/Modified Date/User

| | 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.

8.3.4 Document Archive Settings

 AQUAFOREST SEARCHLIGHT

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

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

[Document Archive Settings](#)

Archive Rule:

Copy to Archive Folder

Archive Template:

%FILENAME%%TIMESTAMP%.%EXT%

Archive Location:

C:\TestFiles\Archive

Archive source Images to Archive Folder:
Yes

Archive source PDFs to Archive Folder
(even when versioning is on):
Yes


Refresh

Save

Service Status : Running

| 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 PDFs to Archive folder | If enabled, this will Archive the source PDFs to the Archive folder (even when versioning is enabled within SharePoint). |

8.3.5 Aquaforest OCR Settings

 AQUAFOREST SEARCHLIGHT

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

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

OCR Engine:
☒ Aquaforest ☐ Extended (IRIS)

[General Settings](#) [PDF Source Settings](#) [TIFF Source Settings](#)

Auto Rotate
Off

Deskew
Off

Remove Lines
Off

JBIG2 Compression
Off

Despeckle
No Despeckle

OCR Language
English

Box Graphics
Remove Box Lines in OCR Processing

Stamps

Advance Flags

| Setting | Description |
|-------------------------|---|
| General Settings | |
| Auto Rotate | Automatically rotate pages so that text flows left to right |
| Deskew | Straighten the image |
| Remove Lines | Remove lines and boxes during OCR processing to improve recognition – particularly in cases where characters touch lines |
| JBIG2 Compression | This option will compress bitonal images in generated PDFs using JBIG2 compression rather than the default Group 4 compression scheme. This will result in smaller PDF file sizes, at a cost of increasing processing time. |
| Despeckle | Remove specks below the specified pixel size from the image |

| Setting | Description |
|-------------------------|---|
| Box/Graphics Processing | <p>By default, if an area of the document is identified as a graphic area then no OCR processing is run on that area. However, certain documents may include areas or boxes that are identified as "graphic" or "picture" areas but that actually do contain useful text.</p> <p>To ensure that the OCR engine can be forced to process such areas there are two options :</p> <p><i>"Treat all Graphics Areas as Text"</i>. This option will ensure the entire document is processed as text.</p> <p><i>"Remove Box Lines in OCR Processing"</i>. This option is ideal for forms where sometimes boxes around text can cause an area to be identified as graphics. This option removes boxes from the temporary copy of the imaged used by the OCR engine. It does not remove boxes from the final image. Technically, this option removes connected elements with a minimum area (by default 100 pixels).</p> |

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

| Setting | Description |
|------------------------------|--|
| MRC | This enables Mixed Raster Compression which can dramatically reduce the output size of PDFs comprising color scans. |
| Remove Hidden Text | Remove existing hidden text (text that was added as a result of a previous OCR) from the PDF file so that the resulting searchable PDF file does not have two layers of the same text. |
| Remove Visible Text | Whether or not to re-OCR existing visible text. |
| DPI | Sets the DPI of rasterized images. If 'Re-image PDF' is used, these images will be added to the output file. |
| PDF/A | Switch on to make sure the output PDF conforms to the PDF/A standards. |
| PDF/A Version | This determines the PDF/A version of the generated PDF. |
| Image Source Settings | |
| MRC | This enables Mixed Raster Compression which can dramatically reduce the output size of PDFs comprising color scans. |
| 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. |

8.3.6 Extended OCR Settings

| Setting | Description |
|-------------|---|
| Auto Rotate | Detect page orientation and correct if required |
| Deskew | Rotates the image to correct its skew angle. |

| Setting | Description |
|----------------------------|---|
| Remove Dark Borders | Removes the dark surrounding from bitonal, grayscale or color images. The dark surrounding of the image is whitened. Note: The dark border should be touching the edge of the image/page for this to work. |
| Keep Original Image | Yes to keep the original image as it is. No to output the image generated after pre-processing is applied. Note: This only applies when the source document is an image (TIFF, BMP, JPEG, PNG) or 'Re-Image PDF' is used when the source is a PDF document. |
| Despeckle | Removes all the groups of connected pixels with a number of pixels below the parameter. |
| Advanced Despeckle | The size of the speckles to remove. |
| Remove White Pixels | By default, despeckle removes black pixels. If set to true, despeckle will remove white pixels rather than black pixels. |
| Work Depth | This parameter (0 – 255) defines how deeply the OCR engine will analyse a page with 255 being the deepest. For poorer quality documents, higher values can give better recognition results. |
| Remove Blank Pages | Set this to true to remove blank pages from output PDF documents. A value needs to be set for sensitivity (see below). |
| Sensitivity | The sensitivity, from 1 to 100. With a high sensitivity, less blank pages are detected. |
| Interpolate | Whether or not to interpolate. |
| Interpolation Mode | Sets the interpolation mode. |
| Interpolation Value | Interpolates the source image to the given resolution. This value (the target resolution) must be greater than the source image's resolution. |
| Language | Set the language(s) to use for OCR. Note: <ul style="list-style-type: none"> Only a maximum of 8 languages can be selected Only the English language can be used in conjunction with an Asian language |
| PDF Source Settings | |
| Re-Image PDF | Each page of the source PDF is rasterized to an image and appended to a new PDF document. |
| Output PDF Version | This determines the PDF version of the generated PDF. |
| Retain Bookmarks | Retains any bookmarks from the source file in the output PDF document when using 'Re-Image PDF'. |
| Retain Metadata | Retains any metadata from the source file in the output PDF document when using 'Re-Image PDF'. |
| 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. |

| Setting | Description |
|---|--|
| DPI | Sets the DPI of rasterized images. If 'Re-image PDF' is used, these images will be added to the output file. However, applying 'Image Compression' or 'iHQC Compression' may reduce the DPI in the output PDF. |
| Image Compression | Compress color JPEG images in generated PDFs |
| JPEG Quality | This parameter (0 – 255) determines the compression/quality of color JPEG images in generated PDFs. 0 gives the smallest file size whilst 255 gives the best quality. |
| JPEG2000 Compression | Use JPEG 2000 compression |
| Compression Mode | The JPEG 2000 compression mode to use. |
| Compression Value | The value to use for the selected compression mode. |
| iHQC Compression | Apply intelligent High Quality Compression |
| Quality Factor | The IHQC quality factor. |
| Compression Level | The iHQC compression level to be used. Level 1 is the basic compression level. Level 3 is the most advanced intelligent High Quality Compression mode. |
| Image Source Settings | |
| Output PDF Version | This determines the PDF version of the generated PDF. |
| Image Compression | Compress color JPEG images in generated PDFs |
| JPEG Quality | This parameter (0 – 255) determines the compression/quality of color JPEG images in generated PDFs. 0 gives the smallest file size whilst 255 gives the best quality. |
| JPEG2000 Compression | Use JPEG 2000 compression |
| Compression Mode | The JPEG 2000 compression mode to use. |
| Compression Value | The value to use for the selected compression mode. |
| iHQC Compression | Apply intelligent High Quality Compression |
| Quality Factor | The IHQC quality factor. |
| Compression Level | The iHQC compression level to be used. Level 1 is the basic compression level. Level 3 is the most advanced intelligent High Quality Compression mode. |
| Advanced Pre-processing Settings | |
| Remove Lines | Whether or not to remove lines from an image (The image must be black and white). |
| Horizontal Clean X | The parameter for cleaning noisy pixels attached to the horizontal lines. |
| Horizontal Clean Y | The parameter for cleaning noisy pixels attached to the horizontal lines. |
| 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. |

| Setting | Description |
|--------------------------|--|
| 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). |

8.3.7 Run Details

| AQUAFOREST SEARCHLIGHT | | | | | | | |
|------------------------|----------|----------------------|-----------------|--|----------------------|-----------------|--|
| Dashboard | | | | Library Settings Help & Support | | | |
| Library Drilldown | | | | Library Settings Document Settings OCR Details Run Details Scheduler Settings Alert Se | | | |
| RUN DATE | AUDIT | | | CONVERSION | | | |
| | Status ▼ | Successful Documents | Error Documents | Status ▼ | Successful Documents | Error Documents | |
| 05/06/2013 15:47:16 | Stopped | 272 | 23 | Not Yet Converted | 0 | 0 | |
| 05/06/2013 16:07:54 | Stopped | 0 | 0 | Stopped | 21 | 2 | |

| DOCUMENT PATH | AUDIT STATUS ▼ | FILE TYPE ▼ | LAST MODIFIED ▼ | PAGES | CONVERSION STATUS ▼ |
|---|----------------|-------------|------------------------|-------|---------------------|
| C:\qa\slqa\Bad\5844bbb2-9c57-11dd-83db-9d068457f227.pdf | error | .PDF | 6/15/2009 11:09:35 AM | 0 | Not Yet Converted |
| C:\qa\slqa\Bad\aaasdad.pdf | error | .PDF | 4/29/2009 9:59:46 AM | 0 | Not Yet Converted |
| C:\qa\slqa\Bad\ABAKHOUR .pdf | error | .PDF | 10/11/2007 12:51:16 PM | 0 | Not Yet Converted |
| C:\qa\slqa\Bad\ABI-LOTF.pdf | error | .PDF | 10/11/2007 12:51:16 PM | 0 | Not Yet Converted |
| C:\qa\slqa\Bad\birneywimage2.pdf | error | .PDF | 10/11/2007 1:05:47 PM | 0 | Not Yet Converted |
| C:\qa\slqa\Bad\example.pdf | error | .PDF | 10/11/2007 1:06:52 PM | 0 | Not Yet Converted |
| C:\qa\slqa\Bad\Fax00000000_1.pdf | error | .PDF | 8/28/2009 5:08:30 PM | 0 | Not Yet Converted |
| C:\qa\slqa\Bad\Functioneel Ontwerp Sharepoint 2003.pdf | error | .PDF | 4/29/2009 9:49:47 AM | 0 | Not Yet Converted |
| C:\qa\slqa\Bad\Legacy-Smith 00678.pdf | error | .PDF | 8/10/2010 1:05:18 AM | 0 | Not Yet Converted |
| C:\qa\slqa\Bad\out_merge.pdf | error | .PDF | 11/4/2008 5:31:51 PM | 0 | Not Yet Converted |
| C:\qa\slqa\Bad\p1000.pdf | error | .PDF | 12/22/2008 2:15:24 PM | 0 | Not Yet Converted |

The Run History list box represents a list of the previous runs carried out on the Document Library. Filters can be applied to limit the runs and individual files shown.

The Generate Log Report button is used to generate the log report of the selected run/documents as a pdf file.

The user can right click on any file in the run details data grid to open the file, open the file location and also view the log entry associated to the particular file.

8.3.8 Run Details Context Menu

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

| DOCUMENT PATH | AUDIT STATUS ▼ | FILE TYPE ▼ |
|---|----------------|-------------|
| C:\qa\slqa\Bad\5844bbb2-9c57-11dd-83db-9d068457f227.pdf | error | .PDF |
| C:\qa\slqa\Bad\aaasdad.pdf | error | .PDF |
| C:\qa\slqa\Bad\ABAKHOUR .pdf | error | .PDF |

Open File
Open File Location

8.3.9 Scheduler Settings

Aquaforest Searchlight

Dashboard Library Settings Help & Support

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

☒ Manual

At: 15 : 31

☐ Once per day

At: 15 : 31

☐ Continuous

Every: 1 Hours

Between 00 : 01 And 23 : 59

☐ Run once

On: 12/02/2015

At: 15 : 31

| Setting | Description |
|--------------|--|
| Manual | This means that the document library has to be run manually by clicking on the "Run" button on the dashboard. |
| Once per day | This allows the document library to be scheduled to run at a specified time each day. |
| Continuous | This allows the document library to be scheduled to run periodically between a start time and end time each day. The periods may be minutes, hours, days or months. For example, a document library may be specified to run every 1 hour between 9:00 and 17:00. |
| Run Once | This allows the document library to be scheduled to run only once at a specified time. |

8.3.10 Alert Settings

Dashboard Library Settings Help & Support Demo

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

Send Email on Job Success ☐ No

Send Email on Job Error ☐ No

Email Settings

From Email Address:

To Email Address:

Email Subject:

Email Message: Test Email

Further processing of this document library has been suspended.

| Setting | Description |
|---------------------------|--|
| Send Email On Job Success | If checked, Aquaforest Searchlight will send an email if a document library finishes processing successfully. The email server configuration should be setup – see the email configuration section for more information. |
| Send Email on Error | If checked, Aquaforest Searchlight will send an email if an error occurs with a document library. The email server configuration should be setup – see the email configuration section for more information. |

8.4 Support

The 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.

8.5 Diagnostic Tool

In order to run the diagnostic tool, click on the “Run Diagnostic” button in the “Help & Support” tab as pointed out in the image below. This will initiate the diagnostic wizard which will run various checks to determine if your system meets all the requirements needed to run Aquaforest Searchlight as well as collect information related to a particular document library. All the gathered information will be made available in a zip file which can be sent to support@aquaforest.com for further investigation.

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

Product Details



Aquaforest Searchlight ®

Version 1.10.160224.0

Copyright © 2016 Aquaforest Ltd

www.aquaforest.com

Support Details

Email:

support@aquaforest.com

[> Reference Guide](#)

[> Troubleshooting Guide](#)

[> Release Notes](#)

[> Online Blogs](#)




Run Diagnostic

Service Status : Running

8.6 Settings

8.6.1 License Settings

 AQUAFOREST SEARCHLIGHT

Dashboard Library **Settings** Help & Support

License Email Theme Advanced

License Type: Permanent

Computer Bound: X

Computer Identifier:

Multi-core: ✓


Max Cores: 64

Document Limit: Unlimited

Trial Stamp: X

Features: Aquaforest OCR: ✓ | Extended OCR: ✓ (Asian: ✓; IHQC: ✓) | SharePoint: ✓ | Pingar: X (iFilter: X)


License Key:

 Update

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

8.6.2 Email Settings

The settings screen allows email server information to be defined. This is used to support the "Send Email on Error" functionality defined in the Scheduler Settings. The password is stored in an encrypted form.


AQUAForest SEARCHLIGHT

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

[License](#)
[Email](#)
[Theme](#)
[Advanced](#)

SMTP Server

SMTP Port

Username

Password

Re-enter Password


Refresh

Save

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

8.6.3 Themes

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


AQUAForest SEARCHLIGHT

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

[License](#)
[Email](#)
[Theme](#)
[Advanced](#)

DARK THEMES

Red

Green

Blue

Purple

Orange

Lime

Emerald

Teal

Cyan

Cobalt

Indigo

Violet

Pink

Magenta

Crimson

Amber

Yellow

Brown

Olive

Steel

Mauve

Taupe

Sienna

LIGHT THEMES

Red

Green

Blue

Purple

Orange

Lime

Emerald

Teal

Cyan

Cobalt

Indigo

Violet

Pink

Magenta

Crimson

Amber

Yellow

Brown

Olive

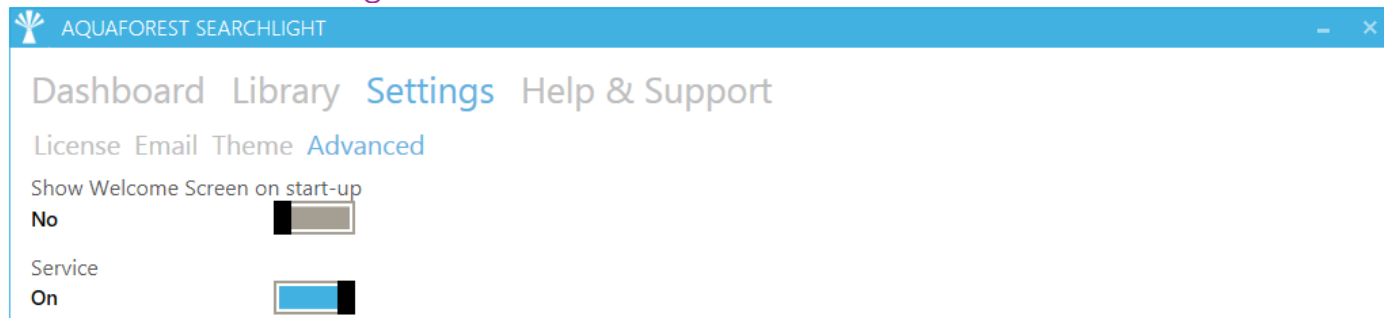
Steel

Mauve

Taupe

Sienna

8.6.4 Advanced Settings



| Setting | Description |
|---------------------------------|---|
| Show Welcome Screen on start-up | Whether or not to show the Welcome Screen when launching the Aquaforest Searchlight UI. |
| Service | Switch to turn the Aquaforest Searchlight service on or off. The service is needed for Audit and OCR. |

9 Aquaforest Searchlight Installation Files

The installation folder (by default C:\Aquaforest\Searchlight) contains the following:

Bin

This folder contains all the executables and libraries that searchlight uses manage the document libraries.

Config

This folder contains the configuration file. This should only be modified on guidance from the support team (support@aquaforest.com).

Data

Searchlight uses a SQLite database file named "Searchlight.db" to store the details, conversions and audit history. We also provide a SQLite Client folder which contains a GUI that you can use to access and view the present state of the database. Note: Modifying the database manually can have a negative impact on Aquaforest Searchlight. It is a good practice to make regular backups of the database (Searchlight.db).

Docs

This folder contains reference guides and release notes for Aquaforest Searchlight.

Extended OCR

This folder contains the executables for the I.R.I.S Extended OCR engine.

License

This folder use to store the license key of searchlight and the license agreement information.

Live

The live folder contains three folders that are used to store the audit and conversion details at runtime.

Livelog

This folder is used to store the live OCR log information of the files that are currently being converted, the log files are grouped into folders by their document library id.

Log

The log folder is used to store the log details of the whole run from auditing to conversion. The log files are grouped into folders by their document library id.

Stats

This folder contains the document library statistics stored in xml files, each document library has an associated xml file with the same name as its id. These file contains a breakdown of searchable, Non searchable documents.

Temp

Before searchlight performs conversions on any pdf document, it copies the files to the temp>source directory and the output is stored in the temp>target directory. After a successful conversion, the file is copied back to its original location.

Tj

This folder contains the executables for the Aquaforest OCR engine.

10 Acknowledgements

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

LEPTONICA

Copyright (C) 2001 Leptonica. All rights reserved.

LIBJPEG

This software is based in part on the work of the Independent JPEG Group.

ZLIB

(C) 1995-2004 Jean-loup Gailly and Mark Adler.

ITEX 4.1.6

Copyright (C) 1999-2009 by Bruno Lowagie and Paulo Soares et al. All Rights Reserved. Binaries distributed under the Mozilla Public License.

CUNEIFORM

Copyright (c) 1993-2008, Cognitive Technologies. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met

Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution. Neither the name of the Cognitive Technologies nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE

LIBTIFF

Copyright (c) 1988-1997 Sam Leffler. Copyright (c) 1991-1997 Silicon Graphics, Inc.

Permission to use, copy, modify, distribute, and sell this software and its documentation for any purpose is hereby granted without fee, provided that (i) the above copyright notices and this permission notice appear in all copies of the software and related documentation, and (ii) the names of Sam Leffler and Silicon Graphics may not be used in any advertising or publicity relating to the software without the specific, prior written permission of Sam Leffler and Silicon Graphics.

THE SOFTWARE IS PROVIDED "AS-IS" AND WITHOUT WARRANTY OF ANY KIND, EXPRESS, IMPLIED OR OTHERWISE, INCLUDING WITHOUT LIMITATION, ANY WARRANTY OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL SAM LEFFLER OR SILICON GRAPHICS BE LIABLE FOR ANY SPECIAL, INCIDENTAL, INDIRECT OR CONSEQUENTIAL DAMAGES OF ANY KIND, OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER OR NOT ADVISED OF THE POSSIBILITY OF DAMAGE, AND ON ANY THEORY OF LIABILITY, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

FREEIMAGE

This software uses the FreeImage open source image library. See <http://freeimage.sourceforge.net> for details. FreeImage is used under the (FIPL), version 1.0.