Contents

1 How to Use this Guide ............................................................................................... 5
2 Technical Support ..................................................................................................... 6
   2.1 Licensing and registration ....................................................................................... 6
   2.2 Contacting Technical Support .................................................................................. 7
3 Introducing contentCrawler ...................................................................................... 8
   3.1 About contentCrawler ............................................................................................ 8
      3.1.1 contentCrawler Framework .................................................................. 8
      3.1.2 contentCrawler OCR Module ................................................................. 8
      3.1.3 contentCrawler Compression Module..................................................... 9
      3.1.4 Additional Processing Modules .............................................................. 9
   3.2 Components of contentCrawler ............................................................................. 10
   3.3 How contentCrawler Framework operates ............................................................... 11
      3.3.1 Prioritization of Services .................................................................... 11
      3.3.2 OCR Module - Non Searchable Content ................................................ 12
      3.3.3 contentCrawler Services .................................................................... 15
      3.3.4 Multi Document Processing ................................................................. 16
      3.3.5 How the DocsCorp contentCrawler Service works ................................. 17
      3.3.6 How contentCrawler Administration Console works ............................... 17
      3.3.7 How the contentCrawler Trial Mode Works (Audit Tool) ......................... 17
4 Planning the Installation ......................................................................................... 19
   4.1 System Requirements .......................................................................................... 19
   4.2 Document Processing Disk Space Requirements ...................................................... 20
   4.3 Microsoft .NET Framework Requirements ................................................................ 21
   4.4 Power Saver Settings ........................................................................................... 21
   4.5 Indexing Software on your Computer ..................................................................... 22
   4.6 Content Repository Requirements .......................................................................... 23
      4.6.1 Autonomy iManage ........................................................................... 23
      4.6.2 File System ..................................................................................... 23
      4.6.3 HP TRIM/Records Manager 8.............................................................. 24
      4.6.4 NetDocuments ................................................................................ 26
      4.6.5 Open Text Livelink 9.7.1 and Content Server 10................................. 27
      4.6.6 Open Text eDocs DM ........................................................................ 27
      4.6.7 ProLaw 11 & 12................................................................................ 27
      4.6.8 SharePoint 2007, 2010, 2013 and Online (O365) .................................. 27
      4.6.9 Worldox GX2 & GX3.......................................................................... 27
   4.7 Optional Requirements ......................................................................................... 29
      4.7.1 Adobe Reader .................................................................................. 29
4.7.2 Microsoft Outlook ................................................................. 29
4.7.3 Lotus Notes ........................................................................ 29
4.8 Firewall Requirements .......................................................... 29
5 Installing contentCrawler .......................................................... 30
5.1.1 Perform the Installation ....................................................... 30
5.1.2 License contentCrawler ........................................................ 31
5.1.3 Additional CPU core Licensing (applies to OCR module only) ...... 32
5.1.4 Check Content Repository Connectivity .................................. 39
5.1.5 Configure Document Cache .................................................. 42
6 Using contentCrawler ............................................................... 43
6.1 File Menu .................................................................................. 43
6.1.1 Save Audit Report ................................................................. 43
6.1.2 New ..................................................................................... 43
6.1.3 About .................................................................................. 44
6.1.4 Help .................................................................................. 44
6.1.5 Options ............................................................................. 45
6.2 Dashboard .................................................................................. 49
6.2.1 Dashboard Features ............................................................... 50
7 contentCrawler OCR Module ..................................................... 53
7.1 Overview .................................................................................. 53
7.2 Creating Text-Searchable Documents ....................................... 53
7.2.1 Image Documents ................................................................. 53
7.2.2 PDF Documents ................................................................. 54
7.2.3 Microsoft Outlook and Lotus Notes Emails ............................ 54
8 contentCrawler Compression Module ......................................... 56
8.1 Overview .................................................................................. 56
9 Backlog and ActiveMonitoring – What is the difference? .......... 57
10 Creating a New Service using the Wizard ................................... 58
10.1.1 Autonomy iManage 8 & 9 ....................................................... 58
10.1.2 File System ......................................................................... 70
10.1.3 HP TRIM/Records Manager 8 ............................................. 82
10.1.4 NetDocuments .................................................................. 94
10.1.5 OpenText Content Server 10 .............................................. 105
10.1.6 OpenText eDocs DM 5 and 10 ........................................... 116
10.1.7 OpenText Livelink ECM 9.7.1 ............................................ 132
10.1.8 ProLaw 11 & 12 ................................................................. 142
10.1.9 SharePoint 2007 ................................................................. 152
10.1.10 SharePoint 2010 and Above ............................................ 165
10.1.11 SharePoint Online ............................................................. 178
10.1.12 Worldox GX2 & GX3 ....................................................... 188
10.2 Compression Advance Settings ............................................. 199
10.2.1 Colour Images ................................................................. 199
10.2.2 Greyscale Images ................................................................. 200
10.2.3 Monochrome Images .......................................................... 201

10.3 Processing steps in Backlog or ActiveMonitoring ..................... 202
10.3.1 Stage 1. Search & Assess .................................................... 202
10.3.2 Stage 2. Process ................................................................. 204
10.3.3 Stage 3. Save .................................................................. 205

10.4 Monitor and Review a service .................................................. 205

10.5 Service Detailed Dashboard .................................................... 206
10.5.1 Detailed Dashboard Status .................................................. 206
10.5.2 Detailed Dashboard Setup Parameters ................................. 207
10.5.3 Detailed Dashboard - Search & Assess ................................. 208
10.5.4 Detailed Dashboard – Process ............................................ 208
10.5.5 Detailed Dashboard – Save ................................................. 209
10.5.6 Live View Reporting Pane .................................................. 210
10.5.7 Report Pane .................................................................. 212

11 Upgrading contentCrawler .......................................................... 216
11.1.1 Backup Internal Database ................................................... 216
11.1.2 Perform the Upgrade ......................................................... 217
11.1.3 Moving contentCrawler Installation ................................. 218

12 Best Practise ............................................................................. 219
12.1 Perform a contentCrawler Server Reboot ................................. 219

13 Worldox – Changing the windows service user password ............... 220
1 How to Use this Guide

This Installation and Administration Guide describes the process of installation and configuration for contentCrawler 3.1.

DocsCorp recommends that you read this Guide in conjunction with the Release Notes. If any conflicts, the Release Notes supersede this and other documentation.

We also recommend that you check the DocsCorp Knowledgebase and Partner or Client Portal for any patches or documentation updates that may have been posted after the initial release of contentCrawler 3.1.
2 Technical Support

2.1 Licensing and registration

contentCrawler is licensed for a specific server/workstation installation. This means that if you have purchased or licensed one copy of the software, you may only install it on one computer for live production work. You are able to install it additionally on other computers for the purposes of testing and evaluating the software, or in the case of computer environments setup for disaster recovery.

When you are provided the contentCrawler installer application, you will also be provided the relevant license key. contentCrawler will automatically install in ‘Trial’ mode however a ‘Full’ product mode license key will be provided once purchased.

The ‘Trial’ mode will allow you to run the audit process to look for documents to be processed in your Content Repository. It will provide statistics on the number of documents requiring processing by contentCrawler. The Trial license will allow you to also see 100 documents that require processing and will allow you to process these 100 documents and save them after processing. Overall statistics for all documents in your Content Repository will be provided, but you will be able to see Document ID’s for 100 documents only.

The ‘Full’ mode enables all functions of contentCrawler to operate fully. Depending on the license key provided, this may either be for a limited timeframe (subscription license) or indefinitely (perpetual license). You will typically be provided with at least 2 license keys linked to your organization’s name.

Company Name:  Your name

This is your organization’s name which must be entered exactly as provided to you with your contentCrawler license keys.

Framework Key:  License Key

This is the license key for the contentCrawler Framework

OCR Module Key:  License Key

This is the license key for the contentCrawler OCR Module.

Compression Module Key:  License Key

This is the license key for the contentCrawler Compression Module.
2.2 Contacting Technical Support

DocsCorp provides technical support for contentCrawler through its global support network. As part of the initial purchase of the software, you are eligible for technical support via phone and email. This will be detailed in the Support and Maintenance agreement provided with your software. The DocsCorp Support group’s primary role is to respond to specific questions on product feature/function, installation, and configuration, as well as to author content for our Web-based Knowledge Base. The DocsCorp Support group works with the DocsCorp Research and Development team to assist with product enhancements and suggestions received from our customers on how the product can be improved.

DocsCorp support offerings include:

- Email support via support@docscorp.com for any technical questions or assistance. This is the preferred method of communicating with the DocsCorp Support team as it allows the team to research your question and provide you with a considered and clear response. When logging support calls via email please provide the following:
  - Your name and email address
  - Daytime contact phone number
  - Your organization name
  - The version of contentCrawler you have installed
  - The modules of contentCrawler installed. E.g. OCR Module
  - Your Operating system version and service pack
  - Your Content Repository product and version
  - Details of the problem
  - Example documents if the issue is document specific.
  - Log files – (Refer to section: 6.1.4 Help)

- Software Updates will be provided to you as part of your Support and Maintenance contract. You will receive notification of any new versions or patches to the contentCrawler software which you are free to download and use while the Support and Maintenance contract is current. Any additional modules for contentCrawler that you have not purchased are only provided to you if you purchase these new modules.

- Telephone support is provided during business hours in the country you purchased your contentCrawler software. Support contact phone numbers are provided on the DocsCorp website at www.docscorp.com
3 Introducing contentCrawler

3.1 About contentCrawler

There are several parts to the contentCrawler application although all are contained in the one physical software installer.

3.1.1 contentCrawler Framework

This is a software application for crawling your Content Repository of documents, which may be a document management system or file system location. Crawling your Content Repository is based on specific search queries, then performing processes on these documents once found, and re-profiling or saving the documents once the processing has been completed. The processes that can be performed vary depending on the contentCrawler process modules you are licensed to use.

3.1.2 contentCrawler OCR Module

The contentCrawler OCR Module provides the ability for the contentCrawler Framework to analyse a set of documents to determine if they are text-searchable, and if not pass them through an OCR process and save them back into your Content Repository.

The contentCrawler OCR Module provides the following functions:

- Analyses PDF documents to determine if they contain text (or if the quantity of text characters found is less than a specific number of characters per page).
- Creates, using OCR technology, a text layer which is applied to any PDF which is not already text-searchable.
- Converts Image documents (TIFF, BMP, JPG and PNG) into text-searchable PDF retaining all the original image content of the document.
- Analyses Microsoft Outlook emails (.MSG) and Lotus Notes emails (.DXL) which contain attachments which are non-text searchable PDF or Image documents and converts those attachments to searchable PDF format. It will also analyse any emails which are themselves attachments to the email and their attachments.

MSG file type support is available for all contentCrawler Content Repository integrations. DXL File type support is only available for the following Content Repository integrations:
- Autonomy iManage
- Open Text eDocs DM
- File System

For further information refer to 8.2.4 Microsoft Outlook and Lotus Notes Emails.

- Can be combined with the compression module to provide end to end multi-function processing in one.
  - Search > Assess > OCR > Compress > Save
3.1.3 contentCrawler Compression Module

The contentCrawler Compress Module provides the ability for the contentCrawler Framework to analyse a set of documents to determine if they are capable of compression, and if so pass them through for compression processing and save them back into your Content Repository. When used in conjunction with the OCR module, documents requiring OCR will be OCR’d and then compressed before saving.

The contentCrawler Compression Module provides the following functions:

- Analyses PDF documents to determine if they contain imagine context and those meeting the image content % are passed for compressing
- Only processed documents for compression when original file sizes greater than minimum size setting
- Converts Image documents (TIFF, BMP, JPG and PNG) into PDF files prior to compressing or to text searchable PDFs when used in conjunction with the OCR module
- Analyses Microsoft Outlook emails (.MSG) and Lotus Notes emails (.DXL) which contain attachments which are PDF or Image documents. Converts and compressed those attachments into the original email. It will also analyse any emails which are themselves attachments to the email and their attachments.

MSG file type support is available for all contentCrawler Content Repository integrations. DXL File type support is only available for the following Content Repository integrations;
  - Autonomy iManage
  - Open Text eDocs DM
  - File System

- Can be combined with the OCR module to provide end to end multi-function processing in one
  - Search > Assess > OCR > Compress > Save

3.1.4 Additional Processing Modules

In the future, additional process modules will be released to perform other functions on your documents in the Content Repository. This will include functions such as converting PDF documents to PDF/A, converting all non-PDF documents to PDF, and so on. These additional modules will fit inside the contentCrawler Framework software application but will be licensed separately from your software supplier and are not supplied as automatic updates as part of your Support and Maintenance agreement.
### 3.2 Components of contentCrawler

The table below describes the main components of contentCrawler.

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DocsCorp contentCrawler Server</td>
<td>This is a windows service that operates regardless of whether the computer is logged in or not, and allows any contentCrawler processes to operate.</td>
</tr>
<tr>
<td></td>
<td>It can be viewed under the Windows Services application and is known as :</td>
</tr>
<tr>
<td></td>
<td>&quot;DocsCorp contentCrawler Server&quot;*</td>
</tr>
<tr>
<td></td>
<td>* Note: Runs under the system context for all environments except Worldox which needs to be run under a user context – see <a href="#">Installation on a Worldox Environment</a></td>
</tr>
<tr>
<td>contentCrawler Administration</td>
<td>This is the application that an administrator can utilize to create, monitor, edit and delete contentCrawler services.</td>
</tr>
<tr>
<td>Console</td>
<td></td>
</tr>
<tr>
<td>Internal Database</td>
<td>contentCrawler retains a database of documents that have been processed so that the administrator can view past activities using the Administrator Console. contentCrawler does not retain full document information – it only retains some document metadata such as file path, Document ID, version, last edit date and size.</td>
</tr>
<tr>
<td></td>
<td>This ensures content of documents are not easily identifiable.</td>
</tr>
</tbody>
</table>
3.3 How contentCrawler Framework operates

contentCrawler is a software application, which is designed to monitor your Content Repository and look for, and process, documents requiring some additional processing after they were originally saved.

A Content Repository could be a Document Management system or simply a series of folders and subfolders of documents in a local or network computer drive.

Often it is important to ensure that standards are met for documents you are retaining in your content repository. At times it can be difficult to enforce these standards for all users for many reasons:

- Users want their documents profiled as quickly as possible
- Many ways in which documents can be profiled (from within applications, via the web, via mobile devices)
- Not easy for a user to determine if a document meets the organization’s standards

contentCrawler is designed to operate while all users in your organization are accessing your content repository in the normal way. It is not necessary, nor is it recommended to allow contentCrawler to only operate out of business hours. contentCrawler accesses all documents through the client software provided by your content repository. For that reason, it is given the same priority to access documents as all other users on your network, so you should not see degradation in performance for your users if they are accessing documents at the same time that contentCrawler is operating.

3.3.1 Prioritization of Services

In many cases, you will have multiple contentCrawler services operating. For example, you may have a backlog service as well as an active monitoring service. contentCrawler automatically understands that all these Services must get an even amount of processor time.

contentCrawler will allow each running service a turn at processing a document. If a service is not waiting to process a document, contentCrawler will move onto the next active Service and allow it to submit a document for processing. So, in this way if you have, for example, 3 Services all running at the same time needing to process documents, each service will submit documents for processing, contentCrawler will process the documents in order and concurrent jobs will be submitted based on your licensed CPU Cores, refer to section 4.3.6 Multi OCR Processing
3.3.2 OCR Module - Non Searchable Content

An excellent example of processing that may be required on documents is to ensure they are text-searchable. Documents may be profiled (saved) by a user without them checking or even knowing whether the document is searchable. For example, if a document is received from a scanner, it may only be an image of the document (without any searchable text). It is not immediately obvious to the user that this document is not searchable so they may profile the document without confirming this and processing the document through an OCR software application to generate the text layer.

contentCrawler Framework working with the contentCrawler OCR Module can look for any documents in the content repository, assess if they are not text searchable, automatically OCR them, and automatically resave the document.

3.3.2.1 How to determine if a PDF is not Searchable

Many PDF documents are created via a process that stores just an image of the document (like a photograph of the page). There is no text information in the document that the user can search for, just millions of dots on a page of various colors and shades that represent an image of the document.

There is no immediately simple way of determining if a PDF document is text-searchable. It can only be done be trial and error.

If you were to open a document that is not text-searchable, you would not be able to use the Find command in Acrobat Reader to find the text – any text you entered in the ‘Find’ field would not be found in the document.

If you open a document in pdfDocs...

...and then try to select text in the document by clicking your mouse on any part of the page, the entire page will be selected.
How an image PDF becomes a text-searchable PDF

PDF documents that contain only images of a page of text are made text-searchable by a process call Optical Character Recognition (OCR). This involves a software application looking at all the dots on a page, and determining what text characters are represented by those dots, including the font type, style and size.

The better the image quality, the more accurate this process will be. For typical typewritten pages that have been scanned, around 99% accuracy of the OCR process should be possible. Handwritten text cannot be understood, unless very clearly written. The OCR process will ignore any graphics on the page it cannot determine to be text. It won’t delete the graphics – it will just fail to place any text behind the image.

It is important to understand that the process of OCRing a document does not in any way affect the image on the document. When you view or print the document after OCRing, it will look exactly as it did previously, with the image remaining with any graphics, pen marks, signatures, etc.

Also, if you have annotated the document with comments, highlighting, etc., these components will remain on the page as before.

The OCR process adds an invisible layer of text BEHIND the image of the text. If you are to open a document that has been OCR’d, and use the ‘Find’ option in pdfDocs you will be selecting the ‘Invisible’ text.

You can then copy/paste using the clipboard any text you select directly into your favourite software application and the text will be visible in that application.

It is important to understand that when you select the text in your PDF, that text may not line up exactly with the image of the text. This is particularly the case where the document has been scanned on a slight angle. In this case, the OCR process will de-skew the image (adjust the image so the text runs horizontally across the page) and determine
the text on the page as if the text is correctly showing running horizontally across the page.

Also, in some cases, the OCR software has to approximate the font size, type and style, and may not have the exact font that the document was created with. In this case also, the text you select or Find may line up exactly with the image of the text, but it will be very close (within a fraction of an inch).

You will find after OCRing a document that it does become slightly larger. Firstly, the actual searchable text will be added to the document, which will be around 5 k per page. Each different font style used in the document will add some additional bytes to the document. This may range in quantity from around 50 kb per document for standard fonts (Arial, Times, etc.) to several MB for unusually complex fonts such as Asian characters.
3.3.3 contentCrawler Services

contentCrawler allows the user to create "contentCrawler Services" which will process sets of documents according to the settings specified by the user, looking for document in a specific location, file type and edit date and assess them for processing.

There are two types of services provided with the contentCrawler.

3.3.3.1 OCR Backlog

This Service is designed to crawl through your existing document store that may have documents saved over many years. It will check all documents that could potentially be processed, will isolate those that meet the processing requirement, from those which do not require processing.

Once the document has been successfully processed, the document will then be saved to the content repository.

For ‘Document Management Systems’, this may be as a new version, revision, attachment, related document or replacing the original document, depending on the features available in your system.

For ‘File Systems’, the document can either be replaced or saved to a new location.

In all cases, where documents are replaced, a backup of the original document will also be created.

This will then ensure over time that all the many thousands or millions of documents in the content repository are searchable and/or compressed.

3.3.3.2 OCR ActiveMonitoring

This Service is designed to watch for any newly saved documents by the users in your organization, and then check to confirm that these documents are searchable. It will perform actions on the document as the Backlog Service, but it will do this for any documents recently saved by any user (for example, in the last hour or so).
3.3.4 Multi Document Processing

Multi document processing is now available in contentCrawler 2 onwards. Multi document processing enables contentCrawler to take full advantage of the physical CPU core processing to run multiple document jobs concurrently across all active services. Previous versions of contentCrawler processed one document at a time during the process stage.

The number of concurrent documents is calculated based on the licensed CPU cores on the machine. By default contentCrawler offers 4 licensed CPU cores, additional CPU cores licenses can be purchased from DocsCorp.

contentCrawler standard (licensed for 4 CPU cores) running on a 4 CPU core machine will run 4 OCR jobs concurrently. However running contentCrawler standard on a 2 CPU core machine will only enable 2 concurrent jobs, it is strongly advised to run contentCrawler on a machine with the same number of physical CPU cores as licensed CPU cores to take full advantage of the multi document processing feature and speed improvements.

It is also important to note that concurrent jobs are shared across ALL Active services. Documents assessed as requiring processing across all services are allocated to one main document processing queue and will be processed in order of when they allocated to the queue, for example contentCrawler running 4 concurrent documents with 2 Active Services (Backlog and Active Monitoring), Active Monitoring shows 3 document in progress at the process stage but the Backlog service show 1 documents processing at the process stage, in total 4 concurrent documents are being utilised.

The table below illustrated how contentCrawler will calculate the number of concurrent documents based on the licensed CPU cores available on the machine.

Example

<table>
<thead>
<tr>
<th>Physical CPU cores</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Licensed Cores</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Standard licensed CPU cores

= number of concurrent documents (shared across all active Services)

3.3.4.1 How many CPU cores is contentCrawler licensed for

By default, contentCrawler is licensed for up to 4 CPU cores, details of how many CPU cores you are licensed for can be found under File > About
If additional CPU core licenses have been purchased, refer to section 6.1.3 Additional CPU core Licensing on how to enable.

If you are unsure on how many physical CPU cores are available on the machine. This information can be found on the machine's system information and can be viewed using Start > run > msinfo32.exe

In the example below, you can see that the machine has 4 physical CPU Core(s).

Please note that CPU cores are different from logical processors; contentCrawler will only look at the CPU core count and will set the concurrent job count in accordance with your CPU core license.

3.3.5 How the DocsCorp contentCrawler Service works

The DocsCorp contentCrawler Service allows the ‘Services’ (Backlog or Active Monitoring) to continue looking for and processing any documents in the background regardless of what software applications are running, or even if the user has logged out of Windows on this computer.

The DocsCorp contentCrawler Service is completely separate to the contentCrawler Administration Console and will continue to process any of the started ‘Services’ whether the Console is currently open or not.

3.3.6 How contentCrawler Administration Console works

The contentCrawler Administration Console is a software application which is used to configure contentCrawler process modules and for the system administrator to be able to review the services running for each process module, exceptions that may have occurred during the process, and to configure new services.

3.3.7 How the contentCrawler Trial Mode Works (Audit Tool)

contentCrawler is also provided to clients in some cases as an ‘Audit’ tool which is referred to in the software as ‘Trial Mode’. contentCrawler will automatically install in ‘Trial Mode’, where no license key is required until a ‘Full license’ is entered to enable the full
functionality. This trial mode allows you to analyse the amount of documents in your content repository that have a particular issue or problem (such as not being text searchable) but without actually correcting the issue. Often this Trial Mode (Audit Tool) is provided at no cost to allow you to identify and confirm that you in fact have an issue that is worthwhile solving.

The Audit Tool also provides the ability to perform a test on a random selection of your documents so that it does not have to analyse every single document in your content repository in order to determine the quantity of documents requiring processing. Typically, contentCrawler will be configured to analyse either of these options (whichever is completed first).

- 10% of all searched documents or
- 48 hours of processing

contentCrawler performs the following steps in order to run the Audit.

1. Search for all documents in the content repository that contain documents that could potentially be a problem. In the case of contentCrawler OCR Module, this involves searching for any PDF documents, image documents, or emails. This search returns to contentCrawler only a list of the Document Paths/IDs/Version numbers of these documents.

2. contentCrawler will randomly select documents to test from the list. Documents are not sorted in any order based on date, size, type, etc. so each document is selected randomly.

3. The selected document will be copied from the content repository to contentCrawler and the content analysed to determine if processing is required. For example, if using the contentCrawler OCR Module, this will test each page of the PDF to determine quantity of text characters per page.

4. The Audit tool will note whether the document needs further processing or not.

Due to the random selection of the documents selected for testing, you will find even for a very large database, the overall percentage of documents requiring processing will be well established after only an hour or so of processing. So, typically at the end of an hour, you will be able to ascertain how many documents in the selected document set will require processing.
# Planning the Installation

## 4.1 System Requirements

The contentCrawler Server requires the following:

<table>
<thead>
<tr>
<th>Category</th>
<th>Requirements</th>
</tr>
</thead>
</table>
| **Processor**             | **Minimum:**  
  - 1 CPU  
  - x86 or x64  
  - 1.6Hz or faster  
  - Expected throughput of 5 to 10 seconds per page (Processing 1 document at a time)  
  **Recommended:**  
  - 4 CPUs dedicated  
  - Multiple CPU cores for faster parallel processing.  
  - Upgrade licenses are available for OCR processing 8, 16 or 32 CPU cores.  
  - Expected throughput of 1 to 3 seconds per page (OCRing 4 documents at a time – 1 document per core)  
  - Upgrading to more than 8 cores will provide expected OCR throughput of ½ to 1 second per page depending on speed of access to documents. |
| **Memory**                | **Minimum:**  
  - 4GB  
  **Recommended:**  
  - 8GB  
  - Additional 1 – 2GB for each additional CPU core over 4.  
  Additional memory may be required if other application services will run on the same system. |
| **Hard Disk**             | **Minimum:**  
  - 100GB - free disk space for program files and typical operation  
  Additional disk space may be required to support large documents or if user wishes to pause processing prior to saving processed documents  
  **Recommended:**  
  - 100GB - free disk space for program files and typical operation (or)  
  - 50GB - free disk space for contentCrawler program files on operating system drive and 50GB free disk space for document cache for held documents.  
  - Additional free space may be required if user chooses to hold for review prior to saving processed documents |
| **Support Operating System** | **Minimum:**  
  - Microsoft® Windows Server® 2012 R2*  
  - Microsoft® Windows Server® 2012*  
  - Microsoft® Windows Server® 2008 R2 with SP1*  
  - Microsoft® Windows Server® 2008 with SP2*  
  **Recommended:**  
  - Microsoft® Windows Server® 2012 R2*  
  - Microsoft® Windows Server® 2008 R2 with SP1*  
  (*) Not supported on Server Core Role. |
Additional Requirements

Minimum:
- Microsoft .NET Framework 4.5 or 4.5.1
  - If your operating system is Server 2012 or higher you may already have the required version of Microsoft .NET Framework, If you do not have .NET 4.5 or 4.5.1 installed you should download the full Microsoft .NET Framework 4.5 or 4.5.1 installer from Microsoft.

Recommended:
- Latest updates for Microsoft .NET Framework, such as Microsoft .NET Framework 4.5.1

Virtualized Environments

Running contentCrawler in a virtual machine (VM) requires specialized skills to set up and manage the virtualized environment. In particular, performance of contentCrawler can be affected by the activity of other VMs running on the same infrastructure, as well as how you configure the contentCrawler VM itself.

DocsCorp supports contentCrawler running on a virtualized environment, such as Vmware ESX or vSphere, Microsoft Hyper-V or Citrix XenServer, but we can only offer support for problems which are unrelated to the environment itself. You will need to understand and be prepared to manage your own virtualised environment if you wish to run contentCrawler on such a platform.

The same operating system, disk space and software requirements specified above are required, however depending on your infrastructure, CPU and memory requirement may differ. Use the information provided above as a guide.

It is good practice to start with the minimum requirement for the required resources, monitor the application performance then increase parameters as required.

If you have any questions or concerns on how Virtualized Environments could impact contentCrawler performance, please contact your DocsCorp representative to discuss further

4.2 Document Processing Disk Space Requirements

contentCrawler requires at least 1 GB disk space for program operation, plus an additional 23MB for every page processed in a multi-page document.

The 40 GB disk space requirement stated in section 4.1 System Requirements comprises of 2.5 GB program files, 1 GB program operation, 4 concurrent jobs and a maximum 400 pages per document limit.

\[2.5 + 1 + (4 \times 400 \times 0.023) = 40.3 \text{ (rounded to 40)} \text{ GB}\]

To allow contentCrawler to process larger documents, the available disk space will need to be increased. For example, to process up to 4x 1000 page documents concurrently, 100 GB disk space will be required.

\[2.5 + 1 + (4 \times 1000 \times 0.023) = 95.5 \text{ (round up to 100 GB)}\]

Additional disk space may be required when using the "Hold" option at any stage of the contentCrawler service process. The amount of required available free disk space may vary greatly. Please refer to section 6.1.5 Configure Document Cache
4.3 Microsoft .NET Framework Requirements

contentCrawler requires Microsoft .NET components installed for the software to install and operate.

- Microsoft .NET Framework 4.5/.Net Framework 4.5.1 (.NET 4.5 is already included in Server 2013)

If you do not have .NET 4.5 or 4.5.1 installed you should download the full Microsoft .NET Framework 4.5 or 4.5.1 installer from Microsoft.


You should refer to Microsoft’s website for any pre-requisites for these modules.

4.4 Power Saver Settings

It is recommended that you configure the power saver settings on your computer so that your computer does not go to sleep while contentCrawler is operating, as this will stop all processing of documents. Specifically it is recommended that you set the following:

- Put the computer to sleep - NEVER
- Turn off hard disk after – NEVER

The settings to switch off your monitor after a period of no activity will not affect the running of contentCrawler, so you are free to set these as you wish.

You will also find that contentCrawler will display a message on the screen if you leave the contentCrawler Administrator application open for a long period of time without any activity. This is so that the screen does not have to keep refreshing – click resume when prompted if you want to see the latest progress status.
4.5 **Indexing Software on your Computer**

It is important that there are no other software applications that will attempt to lock or use the contentCrawler database and cache. In some cases you may have installed indexing products like Windows Search or Google Desktop search (and there are many others). It is important that these products are not installed at all (as it reduces efficiency on this computer), or alternatively you can configure them so that any files associated with contentCrawler are not indexed. The main reason for this requirement is that some indexing products will lock a file while it is being read for the purposes of building an index, and contentCrawler will not expect other applications to lock its files.

You should specifically exclude from your search application the following directories:

- The folder where contentCrawler is installed (Typically c:\program files\DocsCorp\contentCrawler’)
- The location of your document cache if you have configured this to be in a different location using the File – > Options – > Cache option in contentCrawler.
- Windows temp directory – c:\windows\temp
4.6 **Content Repository Requirements**

If you are using contentCrawler to access documents in a Document Management software product, then you must have installed the appropriate client software provided by your software supplier as detailed below.

contentCrawler ensures that all documents retrieved from the document management software are done so via the relevant API toolkit supplied with that DMS. contentCrawler does not ever directly access the database of these systems. This ensures compliance with your security rules and also ensures that contentCrawler does not take priority over other users of the repository.

4.6.1 **Autonomy iManage**

ContentCrawler supports Autonomy iManage 8.2, 8.5 and 9. The iManage client (Desksite or Filesite) must be installed on the computer where contentCrawler is installed. contentCrawler can access an unlimited number of iManage Servers and Databases.

4.6.2 **File System**

4.6.2.1 **File System Format**

contentCrawler has been designed for NT File Systems (NTFS) only. Other file system formats may be compatible. Please contact DocsCorp Support for more information.

4.6.2.2 **Permissions**

User credentials with Full Control permissions must be supplied via the Wizard for contentCrawler to access the content repository and backup locations.

Credentials are only used for the specific task of reading, writing and maintaining permission of files in your content repository.

Note. Full Control permissions are required to the content repository (source), any new locations and backup locations. This allows contentCrawler to save the new OCR’d documents back to the content repository with the same permissions as the original document.

4.6.2.3 **Other 3rd Party Services**

contentCrawler uses several attributes or metadata of the document to identify it uniquely, and to ensure that the source document has not changed whilst it was being processed.

It is possible that you may have other services running on your repository of documents. If these services perform all of the following;

- Modifies the content AND
- Does not rename the document filename or path AND
- Does not change the last modified date;

Then it is important to ensure contentCrawler is NOT run on your repository at the same time as this service.

Note. This requirement is only applicable if the service displays ALL of the behaviours listed above.

Please contact DocsCorp Support for more information.
4.6.3 HP TRIM/Records Manager 8

contentCrawler supports HP TRIM version 7.1, 7.2, 7.3 and RM8.0. The HP TRIM (x86bit) Client must be installed on the computer where contentCrawler is installed. contentCrawler does not support HP TRIM (x64bit) Client. The following warning will appear during installation if contentCrawler is unable to detect the HP TRIM (x86bit) client.

![Unsupported Version]

4.6.3.1 HP TRIM Email Record Handling

ContentCrawler does not support the email file types VMBX or EML, however ContentCrawler does support MSG files saved as record and MSG renditions (Native email format) of VMBX records, that have been created when the option ‘capture native e-mail’ has been enabled within your HP TRIM database.

To check if this option has been enabled on your environment, navigate to the following settings area within HP TRIM:

Tools > HP TRIM Administration > System Options > Record > Capturing native e-mail renditions

Below outlines how each email record type is processed

MSG Records

Due to HP TRIM limitations a MSG record cannot be replaced or a new revision permitted. This means that contentCrawler will always save a processed MSG file as a New Rendition of type ‘Native Email Format’ against the original MSG record. However please note that if an existing Rendition of type ‘Native Email Format’ already exists against the original record, contentCrawler will exception at the save stage as only one rendition of type ‘Native email format’ is permitted by HP TRIM.

MSG renditions (Native email format) of VMBX records

Due to HP TRIM limitations a new revision of a VMBX record is not permitted. This means that when dealing with MSG renditions of VMBX records, contentCrawler will always replace the existing MSG renditions (Native Email Format) with the processed MSG file. Note; contentCrawler will always create a backup of the original MSG Rendition before replacing.

EML Records

From HP TRIM 7.3 onwards Email Records are no longer saved as VMBX format and will be saved as EML. Currently contentCrawler does not support the EML file type format. Existing MSG files in HP TRIM 7.3 will be handled in the same way outlined above. It is possible that future releases of contentCrawler will support the EML file type. Please check with DocsCorp support on when this feature will be available.
4.6.4  NetDocuments

4.6.4.1  NetDocuments prerequisites

ContentCrawler supports NetDocuments 16.1 and all 3 NetDocuments data centres US, EU and AU

For NetDocuments integration contentCrawler uses http calls. Although there are no specific client or server prerequisites required it is recommended that the NetDocuments system requirements are met on the machine to ensure successful connectivity to NetDocuments. Highly recommended that IE8 or above is installed

https://support.netdocuments.com/hc/en-us/articles/205220030-System-Requirements

Note: To establish the connection between contentCrawler and NetDocuments. During the creation of the first service in contentCrawler, NetDocuments will prompt for authorisation to allow contentCrawler access to your repository. Access must be granted to successfully establish and permit integration between contentCrawler and your NetDocuments repository.

Once access has been granted, contentCrawler will be listed under ‘My Integrated Apps’ in NetDocuments under Settings > Manage App Access
IMPORTANT – Revoking access for contentCrawler WILL stop all contentCrawler services processing.

4.6.5 Open Text Livelink 9.7.1 and Content Server 10

The Livelink Web Services must be installed and configured correctly within Content Server 10. To verify successful installation, please connect to your Content Server 10 using the following URL, replacing the server name.

Error! Hyperlink reference not valid.

You should see an overview page describing the Document Management Web Service.

The Content Server or Livelink Indexer Services need to be running for new documents to be successfully found by contentCrawler.

4.6.6 Open Text eDocs DM

There are two options available for eDocs DM connectivity. You may either install the Server API or DM Extensions, which also installs the Server API (PCDClient) – both products provide the components necessary for contentCrawler to connect to your Server.

contentCrawler supports from version 5.1 to version 5.3.1 and 10 of the eDocs DM software. Due to limitations in the eDocs connectivity toolkit, contentCrawler can only connect to one Server at a time. That is, you cannot configure contentCrawler on a particular machine to have processes that simultaneously are connecting to two different servers at the same time; however you can setup multiple Services to different libraries.

4.6.7 ProLaw 11 & 12

ContentCrawler supports ProLaw 11 and 12. The ProLaw Client must be installed on the computer where contentCrawler is installed.

4.6.8 SharePoint 2007, 2010, 2013 and Online (O365)

For Microsoft SharePoint 2007, 2010, 2013 and Online (O365) no client or server prerequisites are required, WebServices/Client Object Model are used for connectivity respectively. It is highly recommended not to install contentCrawler on the SharePoint Server.

Is it highly recommended to read the information on how SharePoint versioning configuration can impact on contentCrawler save workflows

- contentCrawler and SharePoint 2007 Versioning
- contentCrawler and SharePoint 2010 & 2013 Versioning
- contentCrawler and SharePoint Online Versioning

4.6.9 Worldox GX2 & GX3

ContentCrawler supports Worldox GX2 and GX3, both x32Bit and x64Bit. contentCrawler does not support GX. The Worldox Workstation (Mirrored) must be installed on the computer prior to where contentCrawler is installed. The user performing the installation of contentCrawler must have all the required Worldox drives mapped. See section Installing on a Worldox environment for further information.
Worldox Indexer (WDINDEX) needs to be running for new documents to be successfully found by contentCrawler.

Ensure the machine System Variable path contains the correct path to the Worldox programs directory on your environment, this will ensure no issues are encountered when contentCrawler requests Worldox for the list of available ‘Profile Groups’.
4.7 Optional Requirements

To review documents source and result documents via the Administration Console, you may find it useful to install some additional software to view these documents. This not required, but may assist in administering contentCrawler.

4.7.1 Adobe Reader

It is recommended that you have installed Adobe Reader on the computer where you install contentCrawler.

You should have installed version 8 or higher of Adobe Reader.

For most users, the free version of Adobe Reader is sufficient, but if you have installed Acrobat Standard or Acrobat Professional, then you do not need to also install Acrobat Reader. You can download the latest free version of Acrobat Reader from www.adobe.com

Note: Windows 8 include Windows Reader which supports PDF. This can also be used to view PDFs from the Administration Console.

4.7.2 Microsoft Outlook

contentCrawler is able to check and process emails that have been saved into your content repository and it can do this without any version of Microsoft Office or its components being installed on your computer.

However, if you want to open Microsoft Outlook emails and view either the original email and/or the Microsoft Outlook email that is produced after processing by contentCrawler, then you should have Microsoft Outlook 2007 or higher installed.

4.7.3 Lotus Notes

contentCrawler is able to check and process emails that have been saved into your content repository and it can do this without any version of Lotus Notes or its components being installed on your computer.

However, if you want to open Lotus Notes emails and view either the original email and/or the Lotus Notes email that is produced after processing by contentCrawler.

4.8 Firewall Requirements

In some situations, mainly if you have 3rd party firewall software installed, you may notice issues with contentCrawler not responding to any commands when you open the contentCrawler Administration application. This is due to the fact that the application is prevented from communicating with the contentCrawler Server.

In this case you should check the following to ensure they are not blocked:

**IP Address:** 127.0.0.1 otherwise known as localhost

**TCP Port:** 65001
5 Installing contentCrawler

5.1.1 Perform the Installation

Installing contentCrawler is a very easy process. Firstly, ensure that you have all the prerequisites installed defined in Minimum System Requirements section of this document.

You will need local administrator rights to be able to install this application but once installed, you are able to completely logoff from the computer.

It is recommended that you exit out of any other software applications that are running on this computer.

Double-click on the contentCrawler-2-1-x-x.msi application to launch the installer and answer each of the Wizard questions.

Once you have accepted the license agreement, the installer should proceed without the need for much intervention. The location where you choose to install the application should have read/write access.

5.1.1 Information on installing on a Worldox Environment Only

Install contentCrawler as per the above, ensuring all system prerequisites have been met, refer to Minimum System Requirements

After the install location has been selected, the installer will prompt for a username and password. These user credentials are required to allow the contentCrawler Windows Service to run under the user context and gain access to the Worldox Application files and profile groups. To prevent access issues, ensure the user has full administrator rights and all mapped drives to the Worldox profile groups have been configured.

The UserName will be automatically populated with the “Current User” logon details, so it is essential that the installation is performed by the same user which will access Worldox, as the Username field cannot be edited.

Also refer to section Worldox - Changing the windows service password
5.1.2 **Silent and/or Script Deployment**

As this is not an application that will be installed across multiple workstations or network users in your office, there is no facilities provided for scripted deployment using an Installer application.

5.1.2 **License contentCrawler**

contentCrawler requires the user to enter the appropriate license keys so that the software can operate correctly. Generally these keys will be provided when you purchase the software. contentCrawler will automatically install in ‘Trial Mode’, no license key is required. ‘Trial Mode’ enables the user to carry out testing, evaluation and run the Audit Tool. To enable the full contentCrawler functionality a ‘Full Mode’ license key is required to be entered. ([Licensing and registration](#))

To record the Full license key, select

File - > About

![Image](image.png)

Now, choose the ‘Enter Product License Key’ button and enter the license information provided to you. Please note that the license keys are ‘linked’ to the ‘Company Name’, so you must enter the ‘Company Name’ exactly as provided to you with the License Key.

![Image](image.png)

Clicking on the **Update** button will validate the license keys and display the expiry date so you can verify the information was entered correctly.
5.1.3 Additional CPU core Licensing (applies to OCR module only)

contentCrawler offers additional CPU core licensing for faster Multi OCR Processing, Refer to section 4.4.6 Multi Processing

Increased Multi OCR Processing can be activated by an additional license keys provided by DocsCorp.

- Increased number of CPU cores support from the default of 4 to a maximum of 32

5.1.3.1 Manage

The information below steps you through the process of activating and maintaining your additional feature licenses

To manage the additional feature license keys, select the button from the following:

File - > About > Start License Manager

The License Manager screen will be displayed.
5.1.3.2 Activation

The licensing activation takes very little time and is carried out with the help of the Activation Wizard, which is built into the License Manager. The Activation Wizard is used to send the necessary information for Internet license activation or used to enter the activation file you receive via email activation.

The activation information is sent as a code (Installation ID) that is generated on the basis of information about the computer on which the program is being installed. No personal information about the user or computer is used for generating this code and this code cannot be used for identifying the user.

To activate a new feature license select, Activate License...

Enter the 28-digit serial number provided by DocsCorp.

Note this serial number is in addition to your main contentCrawler license key and is provided by DocsCorp on purchasing the additional feature module.

The serial number can be copied and pasted into the first input box to avoid typing errors.

Select Next
The license can be activated in a number of ways:

- **Via the Internet (Fast, recommended)**

  Activation is carried out automatically and takes only a few seconds. An Internet connection is required for this type of activation.

- **By Email**

  The user needs to send an e-mail message generated by the program and containing information required for activation. To ensure a quick reply from the mail robot, do not alter the information in the message body or subject field.

- **By Email from another computer**

  This method is suitable, if your computer does not have an Internet connection or email client. The program will generate an email message containing information required for activation and offer you to copy the message and send from another computer.

  In the case of activation by email, the activation license file received in the response email will need to be added during the activation Wizard.

NOTE: The licensing manager and the licensing service only communicate to the Internet when you explicitly select the ‘via the Internet’ method. This is a once-only connection. Once the license is loaded Internet connectivity is no longer needed.

In this example we will be using the ‘Via the Internet’ option
Select the 'Via the Internet' and Next

Select Next

A progress dialog will appear while the activation process is in session
Once the license key has been successfully activated the below confirmation screen will appear.

If the activation process via the Internet fails or the computer does not have Internet access the 'Via email' or 'email from another computer' methods can be used. This will automatically generate an email which can be sent to activate your license.

Once the license key has been successfully activated you will be returned to the 'License Manager', where the newly-activated license will appear. The Licensing Manager can now be closed.
5.1.3.3 Deactivation

The additional licenses can also be deactivated. The deactivated license can then be activated on another computer. A deactivation of the additional license is required when a move of the contentCrawler software is being performed. The additional license needs to be deactivated before the contentCrawler software is uninstalled, this will allow the license to be moved and re-activated on the new computer.

Deactivation takes very little time and is carried out with the help of a Deactivation wizard which is built into the License Manager.

To deactivate a new feature license, right-click on the relevant license entry in License Manager and select Deactivate.

During the deactivation the Activation license file is deleted. Previous copies of this file cannot be used for activation again.

The deactivation can be performed only via the Internet method. Deactivation is carried out automatically and takes only a few seconds. Once the deactivation is complete, the license can be activated on another computer.
5.1.3.4  Reactivation

The additional licenses can be re-activated on the same or different computer. For example, when moving of contentCrawler to another computer, the additional license will need to be de-activated from the existing computer before uninstalling contentCrawler to allow re-activation of the additional feature license on the new computer.

5.1.3.5  License update

If additional modules are purchased or a new license has been given, an update of your license is required. The license update process is similar to the activation process. The update process is carried out with the help of the Update Wizard and can be performed via the Internet or by email. Once the update is complete, the latest program functionality will become available.

To update a new feature license select, Update License....
5.1.4 Check Content Repository Connectivity

5.1.4.1 Autonomy iManage

There are some simple tests to confirm that the computer operating contentCrawler is able to connect to your Content Repository. Simply launch Desksite or Filesite and ensure that you can connect to at least one database and browse for documents.

contentCrawler will require the Administrator user and password to connect to iManage, so you should test that this connectivity works correctly from Desksite or Filesite. You should not use Trusted Login for contentCrawler iManage access.

There are two reasons that the Administrator user name and password must be used:

- contentCrawler requires access to all documents in the Content Repository regardless of user access rights – the Administrator has full access to all documents;
- The contentCrawler Service will be processing all communication with the server even if you are logged off from your current Windows session. As you will not be logged in, Trusted Login will not be valid, so the Administrator user name and password must be used.

On the same computer, contentCrawler will allow you to create multiple contentCrawler services, with each accessing different databases, or on different servers.

5.1.4.2 File System

contentCrawler will require the FULL CONTROL permissions to access the folders that you wish to process. Ensure that the Windows credentials provided has FULL CONTROL permissions to all documents in the content repository that you wish to be crawled.

contentCrawler will also require FULL CONTROL permissions to the backup location. The backup documents will inherit the permissions of the backup folder.

Note: When using the Save as ‘New Location’, contentCrawler will require FULL CONTROL permissions to the New location. The new documents will inherit the permissions of the new location folder, the original document permissions will NOT be maintained.

5.1.4.3 HP TRIM/RM8

There are some simple tests to confirm that the computer operating contentCrawler is able to connect to your Content Repository. Simply launch HP TRIM or HP TRIM Desktop and ensure that you can connect to at least one dataset and browse for documents.

It is also recommended to ensure that the TRIM/RM program directory path has been referenced in the system variable path. This is to ensure successful connection to the server and datasets. Example C:\Program Files (x86)\Hewlett-Packard\HP TRIM

contentCrawler will require FULL CONTROL permissions to all documents within your HP TRIM datasets. It is recommended that the User credential specified in the contentCrawler service has the following Security and Access Control Levels

- **Security Levels** - hierarchy control - ‘Highest’
- **Security caveats** - privacy control – Either set level to ‘Highest’ or grant access to all caveats
- **Access Control** - specific user level access - set to Administrator
NetDocuments

There are some simple tests to confirm that the computer operating contentCrawler is able to connect to your netDocuments Repository. Simply launch internet explorer (IE 8 or above is recommended) and navigate to www.netdocuments.com. Select the correct data centre – NA, EU or AU and log in. Once you can connect, browse to at least one of the cabinet to be searched.

contentCrawler will require FULL CONTROL permissions to all the documents within your repository and cabinets which require processing.

Open Text Livelink 9.7.1 and Content Server 10

The Livelink Web Services must be installed and configured correctly within Content Server 10. To verify successful installation, please connect to your Content Server 10 using the following URL, replacing the server name.

Error! Hyperlink reference not valid..

You should see an overview page describing the Document Management Web Service.

If you don’t see the page appear correctly there are a number of steps to be performed to add web services support. Please refer to the Open Text KnowledgeCenter.

Open Text eDocs DM

These tests must be performed before you launch contentCrawler to ensure that your connectivity is operating correctly.

Due to the limitations of the eDocs DM Server API for connectivity, contentCrawler can only be configured to access one eDocs DM Server at the one time.

If you have only installed the Server API (PCDClient), you must launch the DM Connection Wizard by selecting:

Start menu – > Program Files - > Open Text - > DM Connection Wizard.

When you see the screen below, specify the Server to which you wish to connect and complete the Wizard that is provided onscreen to establish connection with the server of your choice. Your eDocs database administrator should be able to provide you the necessary information relevant for your organization.

The Server API uses DCOM (Distributed Component Object Model) to remotely connect to your eDocs DM Server. A successful connection requires the system running contentCrawler to have DCOM configured correctly. If you are unsure, it is recommended to install DM Extensions, verify connectivity using a DM Extensions client (Outlook or Windows Explorer), and then configure contentCrawler.

ProLaw 11 & 12

contentCrawler will require the FULL CONTROL permissions to access the ProLaw document folders that you wish to process. Ensure that the Windows credentials provided has FULL CONTROL permissions to all documents in the content repository that you wish to be crawled.
Ensure that the windows user which will be used to run the contentCrawler Service has full access to your SharePoint Repository. To check the user has correct access, log into the machine and navigate to the SharePoint URL.

5.14.9  *Worldox GX2 & GX3*
Ensure Worldox workstation (mirrored) has been installed prior to installing contentCrawler. Once contentCrawler has been successfully installed, check that the machine System Variable path contains the correct path to the Worldox programs directory on your environment, this will ensure no issues are encountered when contentCrawler requests Worldox for the list of available ‘Profile Groups’.

It is also recommended that the user running the contentCrawler windows service is the same user specified during the OCR Service creation and that all mapped drives and full control to the Worldox directories have been configured.
5.1.5  Configure Document Cache

contentCrawler will temporarily cache some documents that it is processing on your local computer. Typically this cache area is used when processing a document and in normal operations the document is immediately flushed from the local cache once it has been processed and re-profiled to your repository.

If you use the ‘Hold for Review’ option on the ‘OCR’ stage, documents will be retained on your local computer after the process has been run. This gives the operator the ability to review the document results prior to them being re-profiled back to the repository. If you are using this ‘Hold for Review’ option you need to have sufficient space in the Cache to retain these documents.

If you are processing and holding for review tens of thousands of documents, you will obviously require sufficient disk space to hold these locally.

By default, contentCrawler caches these documents in sub-folders where the contentCrawler application is installed - by default in “c:\program files\DocsCorp\contentCrawler\data”. However, you may specify a different location for this cache, by selecting and using Browse button to choose a different file location.

To change the settings for the cache, select ‘File menu – Options - Cache’. Choose the Browse button, and then browse to the folder you wish to locate the cache.

Documents will be stored in this directory

- During OCR and/or Compression
- When documents are Held for Review at the Process stage
- If document exception errors occur at the Process or Save Stages.

As a safeguard contentCrawler will stop all processing if the cache drive disk space has reached a low amount. This will mean that if your cache directory has been set to C:\ and the available free disk space on C:\ drops to 5GB, contentCrawler will stop processing documents and will stop all Active Services. This is to ensure that contentCrawler does not impact on the normal operations and functionality of the machine by using all available disk space.

By default the disk space threshold is set to 5GB. The setting below can be used to increase or decrease the available disk space threshold which needs to be reached for contentCrawler to stop all processing.
6 Using contentCrawler

6.1 File Menu

The file menu in contentCrawler allows you to define some basic settings controlling all functions of contentCrawler, as well as create new services for contentCrawler to run.

6.1.1 Save Audit Report

This menu option will save a report in .csv format of the current services created and how many documents have been processed.

This produces a report that can be opened in excel showing each of the services that have run and the overall statistics on documents requiring processing, this includes the total number of documents, total number of pages and also the overall size of documents require processing.

Example OCR

6.1.2 New

Use this menu option to create new processes.

The configuration of each Service will vary depending on that process type selected, so is not explained further in this section of the manual. Please refer to the sections on service creation:

Creating A New Process Using the Wizard
6.1.3 About

This window provides details regarding the version of the software you are using as well as the ability to record the License Keys for the product.

6.1.3.1 License Keys

This will display the Modules currently licensed for

![License Keys](image)

By clicking on the [Enter Product License Key] button, you can record the license keys for the contentCrawler Framework and the other contentCrawler modules you have purchased.

The [Start License Manager] option is used to enable the Additional CPU Core licensing. Only use this setting if directed by DocsCorp to enable option features. For further information please refer to 6.1.3 Additional CPU Core Licensing

6.1.4 Help

These menu options provide you direct access to the software logs. These logs are used by contentCrawler to record any internal program errors. In some cases where you are having a technical issue with the software, DocsCorp may request you to provide the logs so that the cause of an issue can be investigated.

The logs are not designed to be read by a typical user of contentCrawler. They are designed to be read by DocsCorp software engineers.

6.1.4.1 Open Server Log

This function opens the `contentCrawler.log` file in Notepad. This file logs any issues found with the DocsCorp contentCrawler Server, running as a Windows Service.

This file is typically located in the same folder as your contentCrawler software application was installed, which generally is located at

C:`\program files`\DocsCorp`\contentCrawler` or located for 64 Bit operating systems at

C:`\Program files(x86)`\DocsCorp`\contentCrawler`.
6.1.4.2 *Open Administration Console Log*
This function opens the *contentCrawlerAdmin.log* in Notepad. This logs any errors that occur with your usage of the contentCrawler administration console. This file can be found in `%appdata%\DocsCorp\contentCrawler`.

6.1.4.3 *Zip and Save Logs*
This command will take both the ‘Server’ and ‘Administration Console’ logs, compresses all files into a single .zip file and prompts the user for a location to save the file. The zip will also include a Service_Info.csv file, which will export the configuration details of all services created within contentCrawler. This is useful when your DocsCorp technical support team member requests the log files from you to resolve an issue. Please note that the log files may be in segments as contentCrawler automatically extends the log files when they get too large. For example, your Server Log could split across a number of separate log files.

6.1.5 *Options*
The options window is used to configure some settings for contentCrawler.

6.1.5.1 *Cache*
This window allows you to configure where the contentCrawler Cache is located. For additional information on this function refer to section 6.1.5 Configure Document Cache.

6.1.5.2 *Email Notifications*
This screen provided the ability to enable and configure email notifications to be sent to an email address or addresses via SMTP. Currently the email notifications are:

- Service Statistics
- A Service has stopped
- Exception count per service increments by x

Note: It is not a requirement for Outlook or other email software to be installed on the machine for SMTP email notifications.

6.1.5.2.1 *Configuration*

- **Enable Email notification** – Select to enable and disable to the email notifications. Once enabled ensure the correct details are configured under settings.
• **Settings** – Select this option to configure the STMP connection information and the recipient and sender's email address(es)

• **Server Settings**
  - **SMTP Server FQDN** – Enter the name or IP address of the STMP server (i.e. exchange.com or 192.0.0.3)
  - **Port** – Select the port used for outgoing mail on the SMTP server, the default STMP port is 25, other ports can be used. Select from the dropdown or enter a specific port.
  - **Use SSL** – Enable this option if your email provider requires SSL/TLS authentication
  - **Ignore Certificate Errors**
  - **Timeout** – Default 100 seconds
  - **Username/Password** – Enter the credentials of a user with access to the email server

**Email settings**

- **From** – Enter the email address of the sender
- **To** – Enter the email address of the recipient, multiple recipients addressed can be entered using a semi colon separator (;)
- **Test Email** – Ensure the email settings are correct by using the Test Email option. An Email Sent confirmation prompt will appear when the test is successful.

### 6.1.5.2.2 Notification options

Select from the list of email notifications available

• **Service Statistics**

  This option will send an email containing a breakdown of the current statistics for each of the services configured within contentCrawler. The statistics will show the values that are displayed within the contentCrawler admin console for each service.

  - **Starting from:** - Selected the date in which you would like the service statistics emails notifications to start
  - **Repeat Every X hours** – Select how often you wish to receive the email, by default this is set to 24hrs
A Service has stopped

This option will send an email when a Service has stopped. The service may have stopped manually, due to disk space running low or an unknown error.

Example of a Service has stopped email notification

- Exception count per service increments by
  - Search & Assess stage
  - Process stage
  - Save stage

These options will send an email when the selected stage exceptions count reaches the amount specified and will continue to email each time the count increments by that value.

For example if the Search and Assess stage option is selected and set to 200, an email will be sent for any service when the Search & Assess stage exceptions reach 200 documents and should the exceptions continue to increase, for every 200 document increment an email will be generated.

Example of an Exception email notification
6.1.5.3 **OCR Languages**

This function allows you to configure the languages to be detected on documents during the OCR process, these settings apply to all Services created in contentCrawler. The new language settings will be applied immediately, no restart of the service or services is required.

One or more languages can be selected or de-selecting using the Add and Remove buttons.

Please note that selecting a large number of languages to be detected during the OCR stage may impact on processing times, we would recommend only select the languages which are required.

6.1.5.4 **Schedule**

This function allows you to specify certain times when contentCrawler processes should automatically run. This is useful if you have specific times when system maintenance or backup procedures are run on your network.

If the Schedule option is not enabled, contentCrawler processes run 24/7 or until the defined process has completed or stopped.

If the ‘Enable Scheduler’ option is selected, then you must specify the days and times that all processes in contentCrawler are allowed to run.

This controls ALL the Processes that you have configured. It is not possible to schedule each process to operate at a different time.

In the screen example provided, this indicates that all contentCrawler processes will only run between 6 pm and 5 am, 7 days a week.

6.1.5.5 **Exit**

This menu item closes the contentCrawler Administration Console. Please note that any services that are started will continue to operate normally regardless of whether the Administration Console is open or not. The only way to stop all processing is to either:

- Use the *Pause All Services* option on the main dashboard
- By specifying a specific time using the Schedule function
- By switching off a particular service on the main dashboard
6.2 Dashboard

The contentCrawler Dashboard displays all the processes that have been defined within the contentCrawler Framework. From the Dashboard you can monitor these processes, modify them, and individually start them as required.

contentCrawler will run all Services with equal priority, so that one Process doesn’t take all computer power for itself. You can create an unlimited number of separate services and monitor them all from this one Dashboard.

The dashboard shows in summary view how each process is performing. To see more specific details on its activity, click on the hyperlink for Service name.

Selecting the Pause all Services icon will cause the ‘On/Off’ switch for all Services to set to ‘Off’ so that no processes are operating. Once you have used the Pause All option, you must then individually turn on the Services you require. You cannot turn ‘ON’ all Services at once. This is so that you only activate those Services that you really need and don’t accidentally switch on Services that you do not want to run again.

At the lower right hand corner of the screen is an indicator of whether the DocsCorp contentCrawler Service (Windows Service) is running correctly. If this does not show as ‘Started’, you should check that the ‘DocsCorp contentCrawler Service’ is started by checking your windows services in Windows Start Menu - > Control Panels - > Administrative Tools - > Services.
6.2.1 Dashboard Features

Click on the Service name to zoom in on the detailed reporting for this Service

Click or slide the switch to turn the Service on or off

Click on 'Settings' to rename this service and review the settings for it.

Click on 'Delete' to permanently remove the Service although the database remembers what documents were saved

Click on 'complete' to display a report of completed documents

Click on 'review' to display the detailed reporting for this service

6.2.1.1 contentCrawler Service Details

By clicking on the hyperlink which shows the name of the Service (such as “Backlog – Compression Demo”), you will be able to view the full operations of the service and monitor its progress. The actual information displayed will vary depending on the purpose of the particular function you have selected.

Using the example of a Backlog Service, when you click on the Service name, the full details are displayed as follows. To return to the main dashboard, closing the view of the current Service, just click on the Dashboard tab again.
6.2.2  **Settings**

Click on the 'Settings' hyperlink to display all the settings for this Service that were entered when the Service was first created. You cannot change these settings other than the name of the service.

You can change the 'Hold for Review' settings while a Service is running by viewing the detailed progress of the Service and clicking the **Hold for Review** checkboxes.

For more details about the individual options offered, please refer to the section in this manual explaining the particular contentCrawler Process Modules you are working with.

6.2.3  **Delete**

Click on the 'Delete' hyperlink to remove this Service. This will not affect any documents that have been created or modified by this Service, but it does mean that all reporting on what the service has completed or is currently processing will be lost without recovery. Note that there is no **undo** option for this once you have confirmed with the OK dialog that you wish to delete the Service.

**IMPORTANT:** Deleting a service is permanent and cannot be undone. After a service is deleted the contentCrawler database will retain the IDs of documents completely processed by the service, i.e., where a document has been assessed as not requiring processing or where the processed document has been saved back to the repository. In these scenarios, these documents will not get reprocessed by any subsequent services. Whereas all incomplete documents will be removed from contentCrawler with no effect to the repository and can be reprocessed by any future services.

6.2.4  **Complete/Review**

Click on the 'Complete' hyperlink to directly access the Service details for this contentCrawler service and immediately display a report of all documents that have been processed completely with this service.
In some cases, you will also see a Review hyperlink. This hyperlink will appear if you have configured the Service to Hold before processing. In this case the Review hyperlink will take you directly to the Service details screen and display those documents currently being held.
7 contentCrawler OCR Module

7.1 Overview

The contentCrawler OCR Module provides the ability to analyse a set of documents to determine if they are text-searchable, and if not pass them through an OCR Process and save them back into your Content Repository.

The contentCrawler OCR Module provides the following functions:

- Analyses PDF documents to determine if they contain text (or if the quantity of text characters found is less than a specific number of characters per page).
- Creates, using OCR technology, a text layer which is applied to any PDF which is not already text-searchable.
- Converts Image documents (TIFF, BMP, JPG and PNG) into text-searchable PDF retaining all the original image content of the document.
- Analyses Microsoft Outlook emails (.MSG) and Lotus Notes emails (.DXL) which contain attachments which are non-text searchable PDF or Image documents and converts those attachments to searchable PDF format. It will also analyse any emails which are themselves attachments to the email and their attachments.

MSG file type support is available for all contentCrawler DMS integrations. DXL File type support is only available for the following DMS integration:
- Autonomy iManage
- Open Text eDocs DM
- File System

For further information refer to 8.2.4 Microsoft Outlook and Lotus Notes Emails.

- Can be combined with the Compression module to provide end to end multi-function processing in one.
  - Search > Assess > OCR > Compress > Save

7.2 Creating Text-Searchable Documents

contentCrawler OCR Module is able to convert documents that are not-text searchable into searchable content. The steps involved vary depending on the type of document.

7.2.1 Image Documents

Some document formats do not support the ability to store any text information. TIFF, BMP, JPG and PNG are graphics formats supported by contentCrawler, and may be stored in your Content Repository but cannot be text-searchable. To make these documents text searchable, they must be converted to a different document format that support the recording of both the image information and text information. The preferred document type that supports both is PDF.

When contentCrawler looks for these Image documents, they are firstly converted into PDF format, and then the OCR process evaluates the image and applies to the document an invisible text layer.

This newly created PDF is then saved into your Content Repository, but it is now a PDF, not an image. You should be aware that from now on, when reviewing the latest version of
that document in your Content Repository, you will be opening a PDF document on your
desktop, and not an image.

**Hint for Open Text eDocs DM:** If you are using eDocs, when the document is re-
profiled as a new version, the document type shown on the profile is automatically
changed to the relevant ACROBAT document format. If profiling as a new version, you
cannot open the older version (s) of the document in eDocs. You will need to use the
'Compatible Apps' command in your DM eDocs explorer to open the older version
document using the appropriate image viewing software.

**Hint for File System:** If a document already exists with the same name as the image
file, but with a PDF file extension, contentCrawler will attempt to save the document with a
unique file name. The file name convention is [original file without ext]-[number].pdf.

**Hint for SharePoint:** If a document already exists with the same name as the image file,
but with a PDF file extension, contentCrawler will attempt to save the document with a
unique file name. The file name convention is [original file without ext]-[number].pdf.

Also read information on how Images are handled due to SharePoint functionality

- contentCrawler and SharePoint 2007 Versioning
- contentCrawler and SharePoint 2010 & 2013 Versioning

### 7.2.2 PDF Documents

Many PDF documents contain an image and no text information. In this case,
contentCrawler will OCR the document and applies a text layer to the PDF. For further
information on how this process is performed, please refer to How an image PDF becomes
a text-searchable PDF in this manual for information on how a PDF document has a text
layer added to it.

In some cases you may have PDF documents (either profiled separately or as attachments
to emails) that have non-searchable content but also have text. For example, you may
have a scanned page from a photocopier which has had a page number applied to it. In
this example the document does have text (the page number) but certainly the rest of the
document is not text-searchable. To cover these types of documents, you can specify the
minimum average number of characters per page which must be found by the
contentCrawler assessing tool for the document to be classified as text-searchable. This is
configured during the creation of the OCR Process; the default value is an average of 40
characters per page. The assessing tool will analyse the document to calculate the average
characters per page, this is done by obtaining the total number of characters within the
whole document and divide this by the number of pages. If the average character count
per page exceeds 40, the document will be classified as not requiring OCR'ing. Example if
the document has 10 pages, and you set this limit as 40 characters, that means there
must be more than 400 characters in the whole document to be assessed as not requiring
OCR'ing.

### 7.2.3 Microsoft Outlook and Lotus Notes Emails

Most Content Repository indexing systems will search for text not only in the body of a
Microsoft Outlook or Lotus Notes email, but also in attachments to those emails.
However, these indexing systems will not be able to search for text in those attachments
unless these attachments are searchable. So, if the attachments are image-only PDFs, TIFFs, etc, then these will not be able to be located.

contentCrawler will interrogate any Microsoft Outlook or Lotus Notes emails and investigate any attachments to those emails. It is important to understand how these attachments are evaluated by contentCrawler.

contentCrawler will also investigate Microsoft Outlook or Lotus Notes email which are themselves attachments of an email. This email attachment will then be interrogated for non-searchable content. contentCrawler will only crawl two levels deep in attachments – it will only look at attachments within an email within an email.

It is easiest to explain this process by example:

**Example 1** – Microsoft Outlook or Lotus Notes email with 4 attachments

Assume you have an email with 4 attachments

- Document A: 1 PDF document containing Image and text information (already searchable).
- Document B: 1 PDF document which is an Image file only (no text)
- Document C: A Microsoft Word Document containing text

contentCrawler will process this email and create an exact copy of the original email (all the text in the email, metadata, sender details, etc remaining as before) but with the attachments swapped for searchable attachments as follows:

- Document A: No change to this attachment
- Document B: This attachment is replaced by the same PDF with a text layer applied to the document so that it is searchable.
- Document C: No change to this document
- Document D: This attachment replaced by a PDF version of the TIFF image with a text layer applied to the PDF so it is searchable.

**Example 2** - Microsoft Outlook or Lotus Notes email with 1 attachment which is itself an email

Assume you have an email with 1 attachment, but that attachment is itself an email. That email attachment itself contains a TIF image.

In this case the result will be a Microsoft Outlook or Lotus Notes email, with its attachment being an Microsoft Outlook or Lotus Notes email, and the attachment to that email will be a text searchable PDF Document in place of the original TIF image file.
8 contentCrawler Compression Module

8.1 Overview

The contentCrawler Compress Module provides the ability for the contentCrawler Framework to analyse a set of documents to determine if they contain Images and capable of compression, and if so pass them through for compression processing and save them back into your Content Repository. When used in conjunction with the OCR module, documents requiring OCR will be OCR’d and then compressed before saving.

The contentCrawler Compression Module provides the following functions:

- Analyses PDF documents to determine if they contain imagine context and those meeting the image content % are passed for compressing.
- Only processed documents for compression when original file sizes greater than minimum size setting.
- Converts Image documents (TIFF, BMP, JPG and PNG) into PDF files prior to compressing or to text searchable PDFs when used in conjunction with the OCR module.
- Analyses Microsoft Outlook emails (.MSG) and Lotus Notes emails (.DXL) which contain attachments which are PDF or Image documents. Converts and compressed those attachments into the original email. It will also analyse any emails which are themselves attachments to the email and their attachments.

MSG file type support is available for all contentCrawler Content Repository integrations. DXL File type support is only available for the following Content Repository integrations:
  - Autonomy iManage
  - Open Text eDocs DM
  - File System

- Can be combined with the OCR module to provide end to end multi-function processing in one.
  - Search > Assess > OCR > Compress > Save
9 Backlog and ActiveMonitoring – What is the difference?

There are two types of Services available.

- Backlog
- ActiveMonitoring

Both of these services perform the same basic task – they look for documents and assess their content to determine which process type is required. The main difference between them is whether they look ‘forwards’ or ‘backwards’.

**Backlog** processes look either at documents last modified in a particular range of dates, or they look for documents that were last modified less than or equal to a particular date. Typically this will be so that you can process the large number of documents you have existing in your Content Repository that you have been collecting over many years prior to implementing contentCrawler. Backlog looks backwards by date through your Content Repository.

**ActiveMonitoring** processes look FORWARD - looking for any newly profiled documents or documents that have been edited ON or AFTER a particular date. These are typically used to continually monitor your Content Repository for documents that are just now being profiled.

**Backlog Processes**

Look backwards in time through your documents

**ActiveMonitoring**

Look forwards in time to look for newly profiled documents.

When creating these services, or monitoring their progress, the same steps are executed by contentCrawler in terms of how the documents are assessed and processed. The difference is the set of documents that are assessed.

When you use the Wizard to create a new service, the questions asked will be almost identical.
10 Creating a New Service using the Wizard

Please refer to your specific Content Repository sections below

10.1.1 Autonomy iManage 8 & 9

To create a new service, go to the File menu - > New and select ‘New’ as shown against the Backlog function or ActiveMonitoring function

The Service Wizard will then prompt you with a series of questions.

Process Options

- **Process Types** - Select one or more process type to be performed on documents in the service

  *Note:* It is highly recommended that if licensed for OCR and Compression Modules; that combined processing services are created so ensure documents are processed in the correct order, ie OCR’d and then Compressed.
10.1.2 Content Repository Connection Settings

- **Content Repository** - Select the brand of Content Repository (Document Management System) – ‘Autonomy iManage 8 and 9’

- **Server** – Enter the Server name containing your documents.

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access all documents in your iManage content repository.

- **Database** - Once you have entered the Server/User name/password, click on the ‘Get Databases’ button and select the database to connect to. You may only select one database per service.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.1.3 Search Settings

- **Search Dates** – choose the documents you want to assess and process based on the dates they were edited.
  
  **ActiveMonitoring Process**
  
  - **Lasted Edited on or after** – any documents that were edited in any way after the date shown here.
  
  **Backlog Process**
  
  - **Edited on and before** – any documents with a last edited date on or before this date
  
  - **Edited between** – any range of dates
  
- **Document Types** – specify the types of documents you would like to search for and assess for processing. You must select at least one type
  
  - **PDF**
  
  - **Images - BMP, JPG, PNG, TIFF**
  
  - **MSG** (Outlook emails)

  contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable.

  Select the MSG attachment file types to be included in the search
  
  - **PDF**
  
  - **Images - BMP, JPG, PNG, TIFF**
  
  - **DXL** (Lotus Notes Emails)

  contentCrawler will check if these emails have any attachments that require processing (option unchecked by default)

  Select the DXL attachment file types to be included in the search
  
  - **PDF**
  
  - **Images - BMP, JPG, PNG, TIFF**

Click **Next** to proceed to the next step in the Add Service Wizard.
10.1.4  Content Repository Search Settings

- **Include versions declared as records** - any document declared as records will also be included in the search. This option is only available in Backlog services

  **Note:** iManage allows documents to be set as ‘Declared as Record’ which prevents any actions or modification to the document, this includes disabling the option for documents to be ‘replaced’. Currently contentCrawler services allows documents ‘declared as records’ to be included in the search, but Save As ‘replace’ will fail with an exception – Access Denied. Only Save As ‘New Version’ or ‘New Related’ is supported and will successfully create a ‘New Version’ of the original ‘Declared as Record’ document.

  Click *Next* to proceed to the next step in the Add Service Wizard.
Assess Settings

- **OCR Assessment** (Only available when Process Type OCR select for service)

  OCR PDF documents with average characters per page less than or equal to [x] - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCRing. Example if the document has 10 pages, and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR’ing.

- **Compression Assessment** (Only available when Process Type Compress select for service)

  Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image content which is greater than 50% of the overall file size.

  **Note:** Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.
The administrator can opt to ONLY process documents which are marked for OCR (default) or can select “Include documents that do not require OCR”, which will also assess ALL documents including existing text searchable PDF files for compression.

**Note:** This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.6  Process Settings

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio.

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to Compression Advance Settings.

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.
The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.

Administrators may only require documents which have gained 15% file size reduction to be saved back to the DMS. This threshold can be configured using the 'Save compressed documents reduced by at least [x] % of file size'. By default, this is set to 15%. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

**Note:** If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be passed through the process stage and saved back to the Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.7  Content Repository Save Settings

- **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  
  - **Autonomy iManage** – you may choose;
    
    - New Version
    - New Related Document
    - Replace Original

  *Note:* When saving as Replace Original document, the current document when found to be containing non-searchable content, will be replaced with a searchable document. The original document will be placed in the location specified in the ‘Backup Location’.

- **Save As – ‘New Version’**

  ![Content Repository Save Settings](image)

  - **Save Image document type as** - Select the Document type as specified in your Content Repository profile fields. contentCrawler will display those relevant document types which are defined in your Content Repository to have a file extension of .PDF.

  - **Retain ‘Declare as record’ flag of original document** – This option is only available when ‘Include version declared as record’ has been selected in the service search criteria. By selecting this checkbox the new version saved by contentCrawler will be set with the ‘Declare as record’ flag.
• **Save As – ‘New Related Document’**

![Image](image.png)

- **Save Image document type as** - Select the Document type as specified in your Content Repository profile fields. contentCrawler will display those relevant document types which are defined in your Content Repository to have a file extension of .PDF.

- **Set document class to OCR if available** – This option can be select to automatically set the new related document class type to ‘OCR’ if this is configured in the iManage database.

- **Retain ‘Declare as record’ flag of original document** – This option is only available when ‘Include version declared as record’ has been selected in the service search criteria. By selecting this checkbox the new related document saved by contentCrawler will be set with the ‘Declare as record’ flag.
• **Save As – 'Replace Original'**

![Image of contentCrawler interface showing 'Save As' and 'Save Image document type as' options]

- **Save Image document type as** - Select the Document type as specified in your Content Repository profile fields. contentCrawler will display those relevant document types which are defined in your Content Repository to have a file extension of .PDF.

- **Backup Location** – When a document is found by contentCrawler to be non-searchable, the document will be replaced with the processed version of the document.

  **Note:** contentCrawler will create a backup folder structure in the following order:
  
  - Backup folder selected or created during service creation process\Database name\Year of file edit date\Month of file edit date\(e.g. Backup folder \Active\2013\03\)
  - It will use the document’s DMS ID as the name of the document and will suffix the name with the version number of the document being backed up.
    (e.g. 82237-2.tiff)

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format ‘domain name\user name’.

  **Note:** Currently contentCrawler services allows documents ‘declared as records’ to be included in the search, but Save As ‘replace’ will fail with an exception – Access Denied. Only Save As ‘New Version’ or ‘New Related’ is supported and will successfully create a ‘New Version’ or ‘New Related’ of the original ‘Declared as Record’ document.

Click **Next** to proceed to the next step in the Add Service Wizard.
10.1.8 **Service Settings**

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click *Finish* to complete the creation of the Service.
10.1.2 File System

To create a new service, go to the File menu - > New and select ‘New’ as shown against the Backlog function or ActiveMonitoring function.

The Service Wizard will then prompt you with a series of questions.

10.1.2.1 Process Options

- **Process Types** - Select one or more process type to be performed on documents in the service.

  **Note**: It is highly recommended that if licensed for OCR and Compression Modules; that combined processing services are created so ensure documents are processed in the correct order, i.e. OCR’d and then Compressed.
10.1.2.2  **Content Repository Connection Settings**

- **Content Repository** - Select the brand of Content Repository – ‘File System’

![Image of Content Repository Connection Settings page]

- **Location** – Specify the path of the folder location that contains the repository of documents to be searched, either enter the full path of the document folder(s) that you wish to check for non-searchable content or use the ‘Browse’ button to select the folder. The folder location can be a local or network location specifying the drive or folder.

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access all documents in the File System content repository. Should be specified with the domain name in the format ‘domain name\user name’.
• **Manage** Use the manage button to select if there are specific folders that are to be excluded from processing. By default all folders and subfolders are automatically selected. You can choose to ignore certain folders, by unchecking the relevant folder in the tree.

Click *Next* to proceed to the next step in the Add Service Wizard.
10.1.2.3 Search Settings

- **Search Dates** – choose the documents you want to assess and process based on the dates they were edited or created.

**ActiveMonitoring Process**

- **Created or lasted edited on or after** – any documents that were edited in any way after the date shown here.

**Backlog Process**

- **Created or lasted edited on or after** – any documents with a created or last edited date on or before this date

- **Created or lasted edited on or after** – any range of dates

- **Document Types** – specify the types of documents you would like to search for and assess for processing. You must select at least one type

  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF**
  - **MSG (Outlook emails)**

  contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable

  Select the MSG attachment file types to be included in the search

  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF**

- **DXL (Lotus Notes Emails)**

  contentCrawler will check if these emails have any attachments that require processing (option unchecked by default)

  Select the DXL attachment file types to be included in the search

  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF**

Click Next to proceed to the next step in the Add Service Wizard.
10.1.2.4  Assess Settings

- **OCR Assessment** (Only available when Process Type OCR select for service)

  **OCR PDF documents with average characters per page less than or equal to [x]** - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCRing. Example if the document has 10 pages, and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR’ing.

- **Compression Assessment** (Only available when Process Type Compress select for service)

  Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image contain which is greater than 50% of the overall file size.

  **Note:** Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.
The administrator can opt to ONLY process documents which are marked for OCR (default) or can select "Include documents that do not require OCR", which will also assess ALL documents including existing text searchable PDF files for compression.

Note: This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.2.5  Process Settings

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio.

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to Compression Advance Settings.

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.
The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.

![Save compressed documents reduced by at least 15% of file size](image)

Administrators may only require documents which have gained 15% file size reduction to be saved back to the DMS. This threshold can be configured using the 'Save compressed documents reduced by at least [x] % of file size'. By default this is set to 15%. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

*Note:* If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be pass through the process stage and saved back to the Content Repository.

Click *Next* to proceed to the next step in the Add Service Wizard.
10.1.2.6  Content Repository Save Settings

- **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  
  - File System – you may choose;
    
    - New Document
    - Replace Original

  ^Note: When saving as Replace Original document, the current document, when found to be containing non-searchable content, will be replaced with a searchable document. The original document will be placed in the location specified in the ‘Backup Location’.

- **Save As – ‘New Document’**

  ![Content Repository Save Settings](image)

  - **Delete Original** – If selected will permanently delete the original document from the source location, however a backup of the original will be taken and placed in the back up location specified.

  - **New Location** – Specify the full path to the new location in which the new processed documents will be saved. The folder location can be a local or network location specifying the drive or folder. The new location will maintain the same folder structure as the source folder structure, maintaining all sub-folders.

  - **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the new location. Should be specified with the domain name in the format ‘domain name\user name’.

  - **Backup Location** – Note: Only displayed if ‘Delete Original’ selected.
contentCrawler will backup the original document in the backup location prior to replacing it with the text searchable file. The backup location will maintain the same folder structure as the source folder structure, maintaining all sub-folders. The 'Backup' folder becomes a mirror image of the original, but will only have subfolders if required to hold a document.

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format 'domain name\user name'.
• **Save As – 'Replace Original'**

![Image of contentCrawler's OCR Backlog - Add Service Wizard]

- **Backup Location** – When a document is found by contentCrawler to be non-searchable, the document will be replaced with the OCR'd version of the document. contentCrawler will backup the original document in the backup location prior to replacing it with the text searchable file. The backup location will maintain the same folder structure as the source folder structure, maintaining all sub-folders. The 'Backup' folder becomes a mirror image of the original, but will only have subfolders if required to hold a document.

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format 'domain name\user name'.
10.1.2.7 **Service Settings**

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click *Finish* to complete the creation of the Service.
10.1.3 HP TRIM/Records Manager 8

To create a new service, go to the File menu - > New and select ‘New’ as shown against the Backlog function or ActiveMonitoring function.

![Image of ActiveMonitoring and Backlog functions]

The Service Wizard will then prompt you with a series of questions.

### 10.1.3.1 Process Options

- **Process Types** - Select one or more process type to be performed on documents in the service.

  **Note:** It is highly recommended that if licensed for OCR and Compression Modules, that combined processing services are created so ensure documents are processed in the correct order, ie OCR'd and then Compressed.
10.1.3.2  

**Content Repository Connection Settings**

- **Content Repository** - Select the brand of Content Repository (Document Management System) – *‘HP TRIM 7’*

- **Workgroup Server** – Specific Workgroup server name, the name of the computer hosting the workgroup server

- **Port Number** – Specify the port used by the Workgroup Server. This field is automatically configured with the HP TRIM default value 1137 or can be changed to the correct port number of your Workgroup Server.

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access all documents in your HP TRIM content repository.

- **Datasets** - Once you have entered the Workgroup Server/Port/User name/password, click on the ‘Get Datasets’ button and select the dataset to connect to. You may only select one dataset per service.

- **Manage** Use the manage button to select if there are specific ‘Record Types’ in the Dataset that are to be excluded from processing. By default all ‘record types’ automatically selected. You can choose to ignore certain ‘record types’, by unchecking the relevant item.
Click Next to proceed to the next step in the Add Service Wizard.
10.1.3.3 **Search Settings**

- **Search Dates** – choose the documents you want to assess and process based on the dates they were edited.

  **Active Monitoring Process**
  - *Lasted Edited on or after* – any documents that were edited in any way after the date shown here.

  **Backlog Process**
  - *Edited on and before* – any documents with a last edited date on or before this date
  - *Edited between* – any range of dates

- **Document Types** – specify the types of documents you would like to search for and assess for processing. You must select at least one type
  - *PDF*
  - *Images - BMP, JPG, PNG, TIFF*
  - *MSG (Outlook emails)*

  contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable

  Select the MSG attachment file types to be included in the search
  - *PDF*
  - *Images - BMP, JPG, PNG, TIFF*

  Click Next to proceed to the next step in the Add Service Wizard.
• **OCR Assessment** (Only available when Process Type OCR select for service)

**OCR PDF documents with average characters per page less than or equal to \([x]\)** - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCRing. Example if the document has 10 pages, and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR’ing.

• **Compression Assessment** (Only available when Process Type Compress select for service)

Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image content which is greater than 50% of the overall file size.

**Note:** Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.
The administrator can opt to ONLY process documents which are marked for OCR (default) or can select “Include documents that do not require OCR”, which will also assess ALL documents including existing text searchable PDF files for compression.

**Note:** This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.3.5  *Process Settings*

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio.

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to [Compression Advance Settings](#).

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.
The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.

![Save compressed documents reduced by at least \(15\%\) of file size](image)

Administrators may only require document which have gained 15% file size reduction to be saved back to the DMS. This threshold can be configured using the 'Save compressed documents reduced by at least \([x]\) % of file size. By default this is set to 15%. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

**Note:** If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be pass through the process stage and saved back to the Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
**10.1.3.6 Content Repository Save Settings**

**Important**: Read information on how contentCrawler handles Email Records – see 5.6.3.1 – HP TRIM Email Record Handling

- **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  
  - **HP TRIM 7** – you may choose;
    
    - *New Revision*
    - *Replace Original*

  ^**Note:** When saving as Replace Original document, the current document when found to be containing non-searchable content, will be replaced with a searchable document. The original document will be placed in the location specified in the ‘Backup Location’.

- **Save As – ‘New Revision’**

- **Backup Location** - Required when processing MSG files saved as VMBX renditions. When a MSG is found by contentCrawler and contains none text-searchable attachments, the MSG Rendition will be replaced with the OCR’d version of the document and a backup of the original MSG rendition taken.

**Note:** contentCrawler will create a backup folder structure in the following order:

- Backup folder selected or created during service creation process\Dataset ID\Year of file modified date\Month of file modified date\(e.g. Backup folder \45\2013\03)
- It will use the document’s DMS Document ID as the name of the document and will append the name with the revision number of the document being backed up. Important note: Record ID containing characters not supported by windows file system (\ / : * ? " < > | ) will be replaced with an _ (underscore) (e.g. G13_37-2.msg)

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format 'domain name\user name'.

Click **Next** to proceed to the next step in the Add Service Wizard.

- **Save As – ‘Replace Original’**

![Image of OCR Backlog - Add Service Wizard]

- **Backup Location** – When a document is found by contentCrawler to be non-searchable, the document will be replaced with the OCR’d version of the document.

  **Note:** contentCrawler will create a backup folder structure in the following order:

  - Backup folder selected or created during service creation process\Dataset ID\Year of file modified date\Month of file modified date\ (e.g. Backup folder \45\2013\03)

  - It will use the document’s DMS Document ID as the name of the document and will append the name with the revision number of the document being backed up. Important note: Record ID containing characters not supported by
windows file system ( \ / : * ? " < > | ) will be replaced with an _ (underscore) (e.g. G13_37-2.msg)

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format 'domain name\user name'.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.3.7 Service Settings

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click **Finish** to complete the creation of the Service.
10.1.4 NetDocuments

To create a new service, go to the File menu - > New and select ‘New’ as shown against the *Backlog* function or *ActiveMonitoring* function.

The Service Wizard will then prompt you with a series of questions.

10.1.4.1 *Process Options*

- **Process Types** - Select one or more process type to be performed on documents in the service.

  **Note:** It is highly recommended that if licensed for OCR and Compression Modules; that combined processing services are created so ensure documents are processed in the correct order, ie OCR’d and then Compressed.
Content Repository Connection Settings

- **Content Repository** - Select the brand of Content Repository (Document Management System) – *NetDocuments*

- **Location** - You must select your netDocuments datacentre location.

- **Authenticate (Re-direction to NetDocuments login page)** – Select to authenticate to the selected data centre. The following NetDocuments authentication screen will be displayed

- You should specify a full administrator user name and password that provides FULL CONTROL permissions to access all documents in your NetDocuments content Repository.

Click Login to authenticate and proceed to the next step in the Add Service Wizard.
• **Username** – (Read only) Populated after successful authentication to NetDocuments

**IMPORTANT:** Authentication to netdocuments is valid for 12 months. Re-authentication will need to occur before the expiry date shown to ensure the service continues without error. Re-authentication can be performed via the service settings screen. Email notifications can also be configured under File > options to notify when authentication is due to expire.

• **Cabinet(s)** – Once you have successfully authenticated to NetDocuments. A list of your repository cabinets will be pre-populated. You may only select one cabinet per service.

**Note:** Authenticating contentCrawler to NetDocuments for the first time will display the following prompt. You must grant contentCrawler access to your
NetDocuments repository for successful processing. Once granted you will not be prompted with this screen when created subsequent services.

10.1.4.3 Search Settings

**Important:** contentCrawler will only search for the ‘Official version’ of a document

- **Search Dates** – choose the documents you want to assess and process based on the dates they were edited.

**ActiveMonitoring Process**

- **Lasted Edited on or after** – any documents that were edited in any way after the date shown here.

**Backlog Process**

- **Edited on and before** – any documents with a last edited date on or before this date
- **Edited between** – any range of dates

- **Document Types** – specify the types of documents you would like to search for and assess for processing. You must select at least one type

  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF**
  - **MSG (Outlook emails)**

  contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable

  Select the MSG attachment file types to be included in the search

  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF**

  **IMPORTANT:** Version 1 of a MSG file cannot be replace. ContentCrawler will process version 1 of a MSG file and save as a new official version.

Click Next to proceed to the next step in the Add Service Wizard.
10.14.4  Assess Settings

- **OCR Assessment** (Only available when Process Type OCR select for service)

  **OCR PDF documents with average characters per page less than or equal to [x]** - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCR’ing. Example if the document has 10 pages, and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR’ing.

- **Compression Assessment** (Only available when Process Type Compress select for service)

  Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image content which is greater than 50% of the overall file size.

  **Note:** Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.
The administrator can opt to ONLY process documents which are marked for OCR (default) or can select “Include documents that do not require OCR”, which will also assess ALL documents including existing text searchable PDF files for compression.

**Note:** This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.4.5  Process Settings

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio.

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to Compression Advance Settings.

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.
The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.

![Save compressed documents reduced by at least \( \frac{15}{\text{of file size}} \)](image)

Administrators may only require document which have gained 15% file size reduction to be saved back to the DMS. This threshold can be configured using the 'Save compressed documents reduced by at least \( [x] \% \) of file size. By default this is set to 15%. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

**Note:** If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be pass through the process stage and saved back to the Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
10.14.6 Content Repository Save Settings

- **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  - **NetDocuments** – you may choose;
    - **New Version**
    - **Replace Original**

  ^Note: When saving as Replace Original document, the current document when found to be containing non-searchable content, will be replaced with a searchable document. The original document will be placed in the location specified in the ‘Backup Location’.

- **Save As – ‘New Version’**

- **Mark new version ‘Locked’ when original version is locked** – By selecting this checkbox the new version saved by contentCrawler will be set as ‘Locked’

  Important: contentCrawler will automatically save the new version as the ‘Official’, this is to ensure that the new OCR content is indexed and available for searching in NetDocuments. Unofficial version are not indexed by NetDocuments.
• **Save As – ‘Replace Original’**

![Image of Content Repository Save Settings dialog box]

• **Backup Location** – When a document is found by contentCrawler to be non-text searchable, the document will be replaced with the processed version of the document.

**Note:** contentCrawler will create a backup folder structure in the following order:

- Backup folder selected or created during service creation process \CabinetID\Year of backup date(YYYY)\Month of backup date(MM)\ (e.g. Backup folder \NG-2RSNMWWP\Active\2013\04)
- It will use the document’s DMS ID as the name of the document and will append with the version number (e.g. 4816-5960-8873-1.tiff)

• **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format ‘domain name\user name’.

• **Mark new version ‘Locked’ when original version is locked** – By selecting this checkbox the new version saved by contentCrawler will be set as ‘Locked’

Click *Next* to proceed to the next step in the Add Service Wizard.
10.1.4.7 **Service Settings**

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click **Finish** to complete the creation of the Service.
10.1.5 OpenText Content Server 10

To create a new service, go to the File menu - > New and select ‘New’ as shown against the Backlog function or ActiveMonitoring function.

The Service Wizard will then prompt you with a series of questions.

10.1.5.1 Process Options

- **Process Types** - Select one or more process type to be performed on documents in the service.

  **Note:** It is highly recommended that if licensed for OCR and Compression Modules; that combined processing services are created so ensure documents are processed in the correct order, ie OCR’d and then Compressed.
10.1.5.2 *Content Repository Connection Settings*

- **Content Repository** - Select the brand of Content Repository (Document Management System) – ‘Open Text Content Server 10’

![Content Repository Connection Settings](image)

- **Service URL** - You must specify a valid Server URL.  
  Note: Different URLs are required for specific versions of Content Server
  - Content Server 10 – http://{servername}/les - services
  - Content Server 10.5 and above - http://{servername}/cws
  - Content Server 10/5 and above (OTDS) - http://{servername}:port

  contentCrawler does not currently support https configured environments

- **Authentication** – Select the authentication method required

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access all documents in your Open Text Content Server 10 content Repository.

  Click Next to proceed to the next step in the Add Service Wizard.
10.1.5.3  Content Repository Search Settings

- **Search Slice** – Select the search slice defined in your Content Server repository. Default configurations will be set to LES Enterprise.

![Content Repository Search Settings](image)

- **Location ID (Optional)** – Specify a folder or workspace location ID to further refine the search location.

- **Additional Query Conditions (Optional)** – Specify additional conditions to further refine the search criteria. For Example: ("OTDCategory":"47551")

Click *Next* to proceed to the next step in the Add Service Wizard.
10.1.5.4  Search Settings

- **Search Dates** – choose the documents you want to assess and process based on the dates they were edited.

**Active Monitoring Process**
- **Lasted Edited on or after** – any documents that were edited in any way after the date shown here.

**Backlog Process**
- **Edited on and before** – any documents with a last edited date on or before this date
- **Edited between** – any range of dates

- **Document Types** – specify the types of documents you would like to search for and assess for processing. You must select at least one type
  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF**
  - **MSG (Outlook emails)**

  contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable

Select the MSG attachment file types to be included in the search
  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF**

Click Next to proceed to the next step in the Add Service Wizard.
10.1.5.5  Assess Settings

- **OCR Assessment** (Only available when Process Type OCR select for service)

  **OCR PDF documents with average characters per page less than or equal to [x]** - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCRing. Example if the document has 10 pages, and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR’ing.

- **Compression Assessment** (Only available when Process Type Compress select for service)

  Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image content which is greater than 50% of the overall file size.

  **Note:** Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.
The administrator can opt to ONLY process documents which are marked for OCR (default) or can select “Include documents that do not require OCR”, which will also assess ALL documents including existing text searchable PDF files for compression.

**Note:** This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click Next to proceed to the next step in the Add Service Wizard.
**Process Settings**

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio.

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to [Compression Advance Settings](#).

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.
The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.

![Save compressed documents reduced by at least 15% of file size](image)

Administrators may only require document which have gained 15% file size reduction to be saved back to the DMS. This threshold can be configured using the 'Save compressed documents reduced by at least [x] % of file size. By default this is set to 15%. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

**Note:** If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be pass through the process stage and saved back to the Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.5.7  Content Repository Save Settings

- **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  - **Open Text Content Server 10** – you may choose;
    - *New Version*
    - *New Rendition*

- **Save As – ‘New Version’**

![Content Repository Save Settings](image)

- **Set creator to contentCrawler credentials** – When enabled the creator on the new version saved by contentCrawler will automatically be set to the user credentials specified in your contentCrawler service. If unchecked, the original document creator will be retained.

  **Important** - Only available Save As New Version when using Content Server authentication. This option is not currently supported for Windows Authentication or OTDS authentication and will retain the original creator on new versions.
• **Save As – ’New Rendition’**

Click *Next* to proceed to the next step in the Add Service Wizard.
10.1.5.8 **Service Settings**

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click *Finish* to complete the creation of the Service.
10.1.6 OpenText eDocs DM 5 and 10

To create a new service, go to the File menu - > New and select 'New' as shown against the Backlog function or ActiveMonitoring function.

The Service Wizard will then prompt you with a series of questions.

10.1.6.1 Process Options

- **Process Types** - Select one or more process type to be performed on documents in the service.

  **Note:** It is highly recommended that if licensed for OCR and Compression Modules, that combined processing services are created so ensure documents are processed in the correct order, i.e. OCR'd and then Compressed.
10.1.6.2  Content Repository Connection Settings

- **Content Repository** - Select the brand of Content Repository (Document Management System) – *Open Text eDocs DM*

  ![Content Repository Connection Settings](image)

  *Note:* you must specify the eDocs Server to connect to using the DM Extensions client or configure it as per instructions in this document, section 5.6.5 Open Text eDocs DM

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access all documents in your Content Repository.

- **Library** - You must specify a valid Library name.

  Click *Next* to proceed to the next step in the Add Service Wizard.
10.1.6.3 Search Settings

- **Search Dates** – choose the documents you want to assess and process based on the dates they were edited.

  **ActiveMonitoring Process**
  - **Lasted Edited on or after** – any documents that were edited in any way after the date shown here.

  **Backlog Process**
  - **Edited on and before** – any documents with a last edited date on or before this date
  - **Edited between** – any range of dates

- **Document Types** – specify the types of documents you would like to search for and assess for non-searchable content. You must select at least one type
  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF**
  - **MSG (Outlook emails)**
    
    contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable
    
    Select the MSG attachment file types to be included in the search
    - **PDF**
    - **Images - BMP, JPG, PNG, TIFF**
  - **DXL (Lotus Notes Emails)**
    
    contentCrawler will check if these emails have any attachments that require processing (option unchecked by default)
    
    Select the DXL attachment file types to be included in the search
    - **PDF**
    - **Images - BMP, JPG, PNG, TIFF**

Click Next to proceed to the next step in the Add Service Wizard.
Content Repository Search Settings

- **Include read only versions** – only available in backlog service and means that any read only documents will also be included in the search.

- **Include Published Versions** – only available in backlog service and means that any documents marked as 'Published' will also be included in the search.

Click *Next* to proceed to the next step in the Add Service Wizard.
10.1.6.5  

**Assess Settings**

- **OCR Assessment** (Only available when Process Type OCR select for service)

  **OCR PDF documents with average characters per page less than or equal to [x]** - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCRing. Example if the document has 10 pages, and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR’ing.

- **Compression Assessment** (Only available when Process Type Compress select for service)

  Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image content which is greater than 50% of the overall file size.

  **Note:** Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.
The administrator can opt to ONLY process documents which are marked for OCR (default) or can select "Include documents that do not require OCR", which will also assess ALL documents including existing text searchable PDF files for compression.

**Note:** This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.6.6   Process Settings

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio.

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to Compression Advance Settings.

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.
The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.

Administrators may only require document which have gained 15% file size reduction to be saved back to the DMS. This threshold can be configured using the `Save compressed documents reduced by at least [x] % of file size`. By default this is set to 15%. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

  *Note:* If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be pass through the process stage and saved back to the Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
Content Repository Save Settings

- **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  - **Open Text eDocs DM 5** – you may choose;
    - Attachment
    - New Version
    - New Related Document
    - Replace Original

  *Note:* When saving as Replace Original document, the current document, when found to be containing non-searchable content, will be replaced with a searchable document. The original document will be placed in the location specified in the ‘Backup Location’.

- **Save As – ‘Attachment’**

![Content Repository Save Settings](image)

- **New:** Generate new attachment with Label ‘OCR’ – The attachment label must be 3 characters, starting with an alphabetic character (A-Z)

  - **If attachment existing with label above:**
    - **New:** generate unique attachment label starting at AAA
    - **Replace:** Replace with new attachment
    - **Ignore:** Do not replace with new attachment

- **New:** Generate unique attachment label starting at AAA

- **Change Image file Application type to** - Select the Document type as specified in your Content Repository profile fields. contentCrawler will display those relevant document types which are defined in your Content Repository to have a file extension of .PDF.
• **Save As – ‘New Version’**

![Save As - New Version Image](image)

- **Change Image file Application type to** - Select the Document type as specified in your Content Repository profile fields. contentCrawler will display those relevant document types which are defined in your Content Repository to have a file extension of .PDF.

• **Save As – ‘New Related Document’**

![Save As - New Related Document Image](image)
• **Save As – 'Replace Original**

![OCR Backlog - Add Service Wizard](image)

- **Change Image file Application type to** - Select the Document type as specified in your Content Repository profile fields. contentCrawler will display those relevant document types which are defined in your Content Repository to have a file extension of .PDF.

- **Backup Location** – When a document is found by contentCrawler to be non-searchable, the document will be replaced with the OCR’d version of the document.

  **Note:** contentCrawler will create a backup folder structure in the following order:
  - Backup folder selected or created during service creation process\Library\Year of file edit date\Month of file edit date\ (e.g. Backup folder \Live\2013\03)
  - It will use the document’s DMS ID as the name of the document and will suffix the name with the version number of the document being backed up. (e.g. 82237-2.tiff)

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format ‘domain name\user name’.

Click Next to proceed to the next step in the Add Service Wizard.
### Description:
IMAGE file types stored in OpenText eDocs DM which are converted by contentCrawler to PDF using the Save As ‘Replace Original’ option will retain the original IMAGE file extension in the database table and physical file in the eDocs DM document storage location.

### ContentCrawler Versions
All

### OpenText eDocs DM Versions
All

### Scenario:
Affects contentCrawler Backlog and ActiveMonitoring services set to Save As ‘Replace Original’ where the original document is an IMAGE file types (BMP, JPEG, JPG, PNG, TIF, TIFF).

IMAGE documents are converted to PDF as part of contentCrawler processing, but the extension in the database table and physical file in the eDocs DM document storage location will remain the original IMAGE type e.g. TIF, JPG etc.

This issue does not affect documents whereby the original file type is .PDF or .MSG

This issue does not affect any other ‘save as’ options – only save as ‘Replace Original’
Symptoms: From the OpenText eDocs DM right click menu selecting to email or export the saved OCR’d PDF results in a file with the incorrect file type extension (the original IMAGE file type extension) and the document cannot be opened.

This does not affect Right click open or double click launch workflows as the document profile launch application has been successfully set to the PDF viewing software (default Acrobat) during the contentCrawler save process.

Example

Original JPG document

After processing by contentCrawler Save As ‘Replace Original’, the document profile shows the correct application launch type, which in this example environment has been set to Acrobat.
Selecting to Email or Export the document will result in a PDF document with an incorrect file extension, in this example .JPG

Incorrect extension on email

Incorrect extension on export to desktop

The resulting file, although shown with an image file extension, is a valid .pdf file and can be renamed with the correct file extension which will allow the file to be opened successfully

NOTE: This issue only occurs when saving via contentCrawler as 'Replace original'. This does not occur if the OCR service is set to any of the other Save As types

- New version
- Attachment
- Related Document
**Cause:**
Known limitation of the OpenText eDocs DM API replacing a document will not change the file extension.

contentCrawler uses the OpenText eDocs DM API to change the document type on the profile form, in this example .JPG to .PDF. The document types is changed correctly along with the associated application on the document Profile, however this does not update the underlying OpenText eDocs DM Profile.Components table file extension entry and document storage file extension, although the document type has now been set to pdf and the saved files is a valid .PDF.

After processing and Save As Replace via the API, the document components table and DM storage shows the document with a .JPG file extension

**Impact:**
- Minimum with no loss of data
- Limited workflow affected
  - Right click > Email
  - Right click > Export
- Document is a valid PDF

Client can continue to process as Save As Replace but to be aware of the file extension issue when performing certain workflows

Alternatively should they wish, clients can stop their current contentCrawler and create a new service to save as a different option

- New Version
- Attachment
- Related

Please contact DocsCorp support if further information is required
10.1.6.9  **Service Settings**

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click **Finish** to complete the creation of the Service.
10.1.7 OpenText Livelink ECM 9.7.1

To create a new service, go to the File menu - > New and select ‘New’ as shown against the Backlog function or ActiveMonitoring function

The Service Wizard will then prompt you with a series of questions.

10.1.7.1 Process Options

- **Process Types** - Select one or more process type to be performed on documents in the service

  *Note:* It is highly recommended that if licensed for OCR and Compression Modules; that combined processing services are created so ensure documents are processed in the correct order, ie OCR'd and then Compressed.
10.1.7.2 Content Repository Connection Settings

- **Content Repository** - Select the brand of Content Repository (Document Management System) – ‘OpenText Livellink ECM 9.7.1’

- **Service URL** - You must specify a valid Server URL.

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access all documents in your Open Text LiveLink ECM 9.7.1 Repository.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.7.3  Search Settings

- **Search Dates** – choose the documents you want to assess and process based on the dates they were edited.

  **ActiveMonitoring Process**
  - Lasted Edited on or after – any documents that were edited in any way after the date shown here.

  **Backlog Process**
  - Edited on and before – any documents with a last edited date on or before this date
  - Edited between – any range of dates

- **Document Types** – specify the types of documents you would like to search for and assess for processing. You must select at least one type
  - PDF
  - Images - BMP, JPG, PNG, TIFF
  - MSG (Outlook emails)

  contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable.

  Select the MSG attachment file types to be included in the search
  - PDF
  - Images - BMP, JPG, PNG, TIFF

Click Next to proceed to the next step in the Add Service Wizard.
Assess Settings

- **OCR Assessment** (Only available when Process Type OCR select for service)

  **OCR PDF documents with average characters per page less than or equal to [x]** - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCRing. Example if the document has 10 pages, and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR'ing.

- **Compression Assessment** (Only available when Process Type Compress select for service)

  Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image contain which is greater than 50% of the overall file size.

  **Note:** Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.
The administrator can opt to ONLY process documents which are marked for OCR (default) or can select “Include documents that do not require OCR”, which will also assess ALL documents including existing text searchable PDF files for compression.

**Note:** This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.7.5 Process Settings

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio.

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to Compression Advance Settings.

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.
The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.

![Configuration for compression threshold](image)

Administrators may only require document which have gained 15% file size reduction to be saved back to the DMS. This threshold can be configured using the 'Save compressed documents reduced by at least \[x\] % of file size. By default this is set to 15%. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

**Note:** If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be pass through the process stage and saved back to the Content Repository.

Click *Next* to proceed to the next step in the Add Service Wizard.
10.1.7.6 **Content Repository Save Settings**

- **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  
  - **Open Text LiveLink ECM 9.7.1** you may choose;
    - New Version
    - New Rendition

- **Save As – ‘New Version’**

![Content Repository Save Settings](image)
• **Save As – ‘New Rendition’**

![Image of the Save As window]

Click **Next** to proceed to the next step in the Add Service Wizard.
Service Settings

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click *Finish* to complete the creation of the Service.
10.1.8 ProLaw 11 & 12

To create a new service, go to the File menu -> New and select 'New' as shown against the Backlog function or ActiveMonitoring function

The Service Wizard will then prompt you with a series of questions.

10.1.8.1 Process Options

- Process Types - Select one or more process type to be performed on documents in the service

Note: It is highly recommended that if licensed for OCR and Compression Modules; that combined processing services are created so ensure documents are processed in the correct order, i.e. OCR’d and then Compressed.
10.1.8.2  Content Repository Connection Settings

- **Content Repository** - Select the brand of Content Repository (Document Management System) – *ProLaw*

![Content Repository Connection Settings](image)

- **Auto Configure** - will pre-populate the connection details based on the information found by contentCrawler in the ProLaw.ini, by default this is located in C:\Program Files\ProLaw

*Note:* Some environments may store the ProLaw.ini in a different location, in this scenario contentCrawler will be unable to locate the ProLaw.ini automatically and the Auto Configure option will display a prompt enabling the correct ProLaw.ini to be located via the browse option. contentCrawler will then read the connection information and populate the fields. Alternatively if the connection information is known, these can be entered manually in the fields provided.

- **Test Connection** allows the Administrator to verify that the connection information is correct

Click Next to proceed to the next step in the Add Service Wizard.
10.1.8.3 ProLaw Document Repository Location

- **Location** – Specify the path of the folder location that contains the ProLaw repository of documents to be searched, either enter the full path of the document folder(s) that you wish to check for non-searchable content or use the ‘Browse’ button to select the folder. The folder location can be a local or network location specifying the drive or folder.

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access all documents in the ProLaw content repository. Should be specified with the domain name in the format ‘domain name\user name’.

- **Manage** Use the manage button to select if there are specific folders that are to be excluded from processing. By default all folders and subfolders are automatically selected. You can choose to ignore certain folders, by unchecking the relevant folder in the tree.

Click *Next* to proceed to the next step in the Add Service Wizard.
10.1.8.4  Search Settings

- **Search Dates** – choose the documents you want to assess and process based on the dates they were edited.

  **Active Monitoring Process**

  - Created or lasted edited on or after – any documents that were edited in any way after the date shown here.

  **Backlog Process**

  - Created or lasted edited on or after – any documents with a created or last edited date on or before this date
  - Created or lasted edited on or after – any range of dates

- **Document Types** – specify the types of documents you would like to search for and assess for processing. You must select at least one type

  - PDF
  - Images - BMP, JPG, PNG, TIFF
  - MSG (Outlook emails)

  contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable

  Select the MSG attachment file types to be included in the search
  - PDF
  - Images - BMP, JPG, PNG, TIFF

Click Next to proceed to the next step in the Add Service Wizard.
**OCR Assessment** (Only available when Process Type OCR select for service)

**OCR PDF documents with average characters per page less than or equal to [x]** - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCRing. Example if the document has 10 pages and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR’ing.

**Compression Assessment** (Only available when Process Type Compress select for service)

Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image content which is greater than 50% of the overall file size.

*Note: Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.*
The administrator can opt to ONLY process documents which are marked for OCR (default) or can select “Include documents that do not require OCR”, which will also assess ALL documents including existing text searchable PDF files for compression.

**Note:** This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click Next to proceed to the next step in the Add Service Wizard.
Process Settings

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio.

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to Compression Advance Settings.

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.
The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.

![Save compressed documents reduced by at least % of file size](image)

Administrators may only require document which have gained 15% file size reduction to be saved back to the DMS. This threshold can be configured using the 'Save compressed documents reduced by at least \([x]\) % of file size. By default this is set to 15%. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

**Note:** If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be pass through the process stage and saved back to the Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.8.7 Content Repository Save Settings

• **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  
  o **ProLaw** – you may choose;
    
    • **Replace Original**

  ^Note: When saving as Replace Original document, the current document, when found to be containing non-searchable content, will be replaced with a searchable document. The original document will be placed in the location specified in the 'Backup Location'. The replaced file will remain with the same date created, modified, etc as the original document.

• **Save As** – ‘Replace Original’

![OCR Backlog - Add Service Wizard](image)

• **Backup Location** – When a document is found by contentCrawler to be non-searchable, the document will be replaced with the OCR’d version of the document. contentCrawler will backup the original document in the backup location prior to replacing it with the text searchable file. The backup location will maintain the same folder structure as the source folder structure, maintaining all sub-folders. The ‘Backup’ folder becomes a mirror image of the original, but will only have subfolders if required to hold a document.

• **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format ‘domain name\user name’.
10.1.8.8  Service Settings

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click *Finish* to complete the creation of the Service.
10.1.9 SharePoint 2007

To create a new service, go to the File menu -> New and select 'New' as shown against the *Backlog* function or *ActiveMonitoring* function

The Service Wizard will then prompt you with a series of questions.

10.1.9.1 *Process Options*

- **Process Types** - Select one or more process type to be performed on documents in the service

*Note:* It is highly recommended that if licensed for OCR and Compression Modules; that combined processing services are created so ensure documents are processed in the correct order, ie OCR'd and then Compressed.
10.1.9.2  Content Repository Connection Settings

- **Content Repository** - Select the brand of Content Repository (Document Management System) – ‘SharePoint 2007’

- **Site URL** – Enter the SharePoint 2007 URL

**IMPORTANT Information:** contentCrawler searches SharePoint sites and libraries following the below conditions

- Will search all sites and sub-sites of the specified URL unless excluded using the manage option via the service wizard
- Will only search site libraries of template type ‘Document Library’
- Will NOT search libraries set with the following attributes
  - Hidden
  - Application List
- **Manage** Use the manage button to select if there are specific sub-sites that are to be excluded from processing. By default all sites and sub-sites are automatically selected. You can choose to ignore certain sites and sub-sites, by unchecking the relevant folder in the tree.

- **Authentication** – Windows Authentication or Forms Authentication

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access all document libraries and documents in your SharePoint 2007 Content Repository.

Click *Next* to proceed to the next step in the Add Service Wizard.
10.1.9.3 *Search Settings*

- **Search Dates** – choose the documents you want to assess and process based on the dates they were edited.

**Active Monitoring Process**

- **Lasted Edited on or after** – any documents that were edited in any way after the date shown here.

**Backlog Process**

- **Edited on and before** – any documents with a last edited date on or before this date
- **Edited between** – any range of dates

- **Document Types** – specify the types of documents you would like to search for and assess for processing. You must select at least one type

  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF** – See notes on [image file type handling](#)
  - **MSG (Outlook emails)**

  contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable

  Select the MSG attachment file types to be included in the search

  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF**

Click Next to proceed to the next step in the Add Service Wizard.
10.1.9.4 Content Repository Search Settings

- **Include Minor Versions** – Select to include Minor version in the search (Available in both Active Monitoring and Backlog)

Click *Next* to proceed to the next step in the Add Service Wizard.
• **OCR Assessment** (Only available when Process Type OCR select for service)

**OCR PDF documents with average characters per page less than or equal to [x]** - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCRing. Example if the document has 10 pages, and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR’ing.

• **Compression Assessment** (Only available when Process Type Compress select for service)

Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image content which is greater than 50% of the overall file size.

**Note:** Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.
The administrator can opt to ONLY process documents which are marked for OCR (default) or can select “Include documents that do not require OCR”, which will also assess ALL documents including existing text searchable PDF files for compression.

**Note:** This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click Next to proceed to the next step in the Add Service Wizard.
Process Settings

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to Compression Advance Settings

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.
The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.

![Save compressed documents reduced by at least $15\%$ of file size]

Administrators may only require documents which have gained $15\%$ file size reduction to be saved back to the DMS. This threshold can be configured using the 'Save compressed documents reduced by at least $[x] \%$ of file size'. By default, this is set to $15\%$. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

**Note:** If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be passed through the process stage and saved back to the Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.9.7  Content Repository Save Settings

**Important:** Read information on how SharePoint versioning configuration can impact on contentCrawler save workflows

[contentCrawler and SharePoint 2007 Versioning]

- **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  - **SharePoint 2007** – you may choose;
    - New Version
    - Replace Original

^Note: When saving as Replace Original document, the current document when found to be containing non-searchable content, will be replaced with a searchable document. The original document will be placed in the location specified in the ‘Backup Location’.

- **Save As** – ‘New Version’

![OCR Backlog - Add Service Wizard](image)

- **Backup Location** – Location used by contentCrawler safeguard in scenarios whereby a backup of the original source document will be taken before deletion by SharePoint 2007 if the document version limit has been reached.

**Note:** contentCrawler will create a backup folder structure in the following order:

- Backup folder selected or created during service creation process\file directory reference \Year of file edit date\Month of file edit date\
(e.g. Backup folder \Active\2013\03)

- It will use the document’s DMS ID as the name of the document and will suffix the name with the version number and minor version number of the document being backed up. (e.g. 43-1.1.tiff)

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format ‘domain name\user name’.
- **Save As – 'Replace Original'**

![Image of the Content Repository Save Settings window with 'Save As' set to 'Replace Original']

- **Delete Original Image Version** – This option only applies to image files processed by contentCrawler, if checked the image file, including all previous versions will be deleted. contentCrawler will back-up the image file and all previous versions prior to the deletion.

- **Backup** – Location used by contentCrawler safeguard in scenarios whereby a backup of the original source document will be taken before being replaced by contentCrawler or deletion by SharePoint 2007 if the document version limit has been reached.

  **Note:** contentCrawler will create a backup folder structure in the following order:

  - Backup folder selected or created during service creation process\ file directory reference \Year of file edit date\Month of file edit date\ (e.g. Backup folder \Active\2013\03)
  - It will use the document’s DMS ID as the name of the document and will suffix the name with the version number and minor version number of the document being backed up. (e.g. 43-1.1.tiff)

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format ‘domain name\user name’.

  ![Image of the backup folder structure and credentials input fields]
10.1.9.8  *Service Settings*

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click *Finish* to complete the creation of the Service.
10.1.10 SharePoint 2010 and Above

To create a new service, go to the File menu - > New and select ‘New’ as shown against the Backlog function or ActiveMonitoring function.

The Service Wizard will then prompt you with a series of questions.

10.1.10.1 Process Options

- **Process Types** - Select one or more process type to be performed on documents in the service.

  **Note:** It is highly recommended that if licensed for OCR and Compression Modules; that combined processing services are created so ensure documents are processed in the correct order, ie OCR’d and then Compressed.
10.1.10.2  Content Repository Connection Settings

- **Content Repository** - Select the brand of Content Repository (Document Management System) – *‘SharePoint 2010 and Above’*

- **Site URL** – Enter the SharePoint URL

  **IMPORTANT Information**: contentCrawler searches SharePoint sites and libraries following the below conditions
  
  - Will search all sites and sub-sites of the specified URL unless excluded using the manage option via the service wizard
  - Will only search site libraries of template type ‘Document Library’
  - Will NOT search libraries set with the following attributes
    - Hidden
    - Application List
    - Site Asset Library
• **Manage** Use the manage button to select if there are specific sub-sites that are to be excluded from processing. By default all sites and sub-sites are automatically selected. You can choose to ignore certain sites and sub-sites, by unchecking the relevant folder in the tree.

• **Authentication** – Windows Authentication or Forms Authentication

• **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access all document libraries and documents in your SharePoint 2010/2013 Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.10.3  Search Settings

- Search Dates – choose the documents you want to assess and process based on the dates they were edited.

  ActiveMonitoring Process
  
  o Lasted Edited on or after – any documents that were edited in any way after the date shown here.

  Backlog Process
  
  o Edited on and before – any documents with a last edited date on or before this date
  
  o Edited between – any range of dates

- Document Types – specify the types of documents you would like to search for and assess for processing. You must select at least one type

  o PDF
  
  o Images - BMP, JPG, PNG, TIFF – See notes on image file type handling
  
  o MSG (Outlook emails)

  contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable

  Select the MSG attachment file types to be included in the search

  - PDF
  
  - Images - BMP, JPG, PNG, TIFF

Click Next to proceed to the next step in the Add Service Wizard.
10.1.10.4 Content Repository Search Settings

- **Include Minor Versions** – Select to include Minor version in the search (Available in both Active Monitoring and Backlog)

Click Next to proceed to the next step in the Add Service Wizard.
• **OCR Assessment** (Only available when Process Type OCR select for service)

**OCR PDF documents with average characters per page less than or equal to [x]** - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCRing. Example if the document has 10 pages, and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR’ing.

• **Compression Assessment** (Only available when Process Type Compress select for service)

Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image content which is greater than 50% of the overall file size.

**Note:** Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.
The administrator can opt to ONLY process documents which are marked for OCR (default) or can select “Include documents that do not require OCR”, which will also assess ALL documents including existing text searchable PDF files for compression.

**Note:** This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.10.6  Process Settings

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio.

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to Compression Advance Settings.

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.
The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.

Administrators may only require documents which have gained 15% file size reduction to be saved back to the DMS. This threshold can be configured using the 'Save compressed documents reduced by at least [x] % of file size. By default this is set to 15%. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

**Note:** If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be passed through the process stage and saved back to the Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
**10.1.10.7 Content Repository Save Settings**

**Important:** Read information on how SharePoint versioning configuration can impact on contentCrawler save workflows

contentCrawler and SharePoint 2010 & 2013 Versioning

- **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  - **SharePoint 2010/2013** – you may choose;
    - *New Version*
    - *Replace Original*

  ^Note:** When saving as Replace Original document, the current document when found to be containing non-searchable content, will be replaced with a searchable document. The original document will be placed in the location specified in the ‘Backup Location’.

- **Save As – ‘New Version’**

![OCR Backlog - Add Service Wizard](image)

- **Backup** – Location used by contentCrawler safeguard in scenarios whereby a backup of the original source document will be taken before deletion by SharePoint 2010/2013 if the document version limit has been reached.

  **Note:** contentCrawler will create a backup folder structure in the following order:

  - Backup folder selected or created during service creation process\file directory reference \Year of file edit date\Month of file edit date\(e.g. Backup folder \Active\2013\03)
• It will use the document’s DMS ID as the name of the document and will suffix the name with the version number and minor version number of the document being backed up. (e.g. 43-1.1.tiff)

• **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format ‘domain name\user name’.
• **Save As – 'Replace Original'**

![Image of OCR Backlog - Add Service Wizard](image)

- **Backup** – Location used by contentCrawler safeguard in scenarios whereby a backup of the original source document will be taken before being replaced by contentCrawler or deletion by SharePoint 2010/2013 if the document version limit has been reached.

  **Note:** contentCrawler will create a backup folder structure in the following order:

  - Backup folder selected or created during service creation process\file directory reference \Year of file edit date\Month of file edit date\(e.g. Backup folder Active2013/03)
  - It will use the document’s DMS ID as the name of the document and will suffix the name with the version number and minor version number of the document being backed up. (e.g. 43-1.1.tiff)

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format 'domain name\user name'.
10.1.10.8  Service Settings

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click **Finish** to complete the creation of the Service.
10.1.11 SharePoint Online

To create a new service, go to the File menu - > New and select ‘New’ as shown against the Backlog function or ActiveMonitoring function.

The Service Wizard will then prompt you with a series of questions.

10.1.11.1 Process Options

- **Process Types** - Select one or more process type to be performed on documents in the service.

  **Note:** It is highly recommended that if licensed for OCR and Compression Modules; that combined processing services are created so ensure documents are processed in the correct order, ie OCR’d and then Compressed.
10.11.2 Content Repository Connection Settings

- **Content Repository** - Select the brand of Content Repository (Document Management System) – *SharePoint Online*

- **Site URL** – Enter the SharePoint Online URL

- **Manage** Use the manage button to select if there are specific sub-sites that are to be excluded from processing. By default all sites and sub-sites are automatically selected. You can choose to ignore certain sites and sub-sites, by unchecking the relevant folder in the tree

- **Authentication** – Claim Based Authentication

- **Credentials (Email Address & Password)** - you should specify a full administrator email address and password that provides FULL CONTROL permissions to access all document libraries and documents in your SharePoint Online Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.11.3 Search Settings

- **Search Dates** – choose the documents you want to assess and process based on the dates they were edited.

**Active Monitoring Process**

- **Lasted Edited on or after** – any documents that were edited in any way after the date shown here.

**Backlog Process**

- **Edited on and before** – any documents with a last edited date on or before this date
- **Edited between** – any range of dates

- **Document Types** – specify the types of documents you would like to search for and assess for processing. You must select at least one type
  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF** – See notes on image file type handling
  - **MSG (Outlook emails)**

  contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable

  Select the MSG attachment file types to be included in the search
  - **PDF**
  - **Images - BMP, JPG, PNG, TIFF**

Click Next to proceed to the next step in the Add Service Wizard.
10.11.4  Content Repository Search Settings

- **Include Minor Versions** – Select to include Minor version in the search (Available in both Active Monitoring and Backlog)

Click *Next* to proceed to the next step in the Add Service Wizard.

10.11.5  Assess Settings

- **OCR Assessment** (Only available when Process Type OCR select for service)
OCR PDF documents with average characters per page less than or equal to [x] - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCRing. Example if the document has 10 pages, and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR'ing.

- **Compression Assessment** (Only available when Process Type Compress select for service)

Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image content which is greater than 50% of the overall file size.

*Note: Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.*

![Compression Assessment](image)

The administrator can opt to ONLY process documents which are marked for OCR (default) or can select “Include documents that do not require OCR”, which will also assess ALL documents including existing text searchable PDF files for compression.

*Note: This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability*

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click **Next** to proceed to the next step in the Add Service Wizard.
10.1.11.6 Process Settings

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio.

![Compression Settings](image)

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to Compression Advance Settings.

![Advanced Compression Settings](image)

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.

The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.
Administrators may only require document which have gained 15% file size reduction to be saved back to the DMS. This threshold can be configured using the ‘Save compressed documents reduced by at least \[x\] % of file size. By default this is set to 15%. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

**Note:** If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be pass through the process stage and saved back to the Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
10.11.7 Content Repository Save Settings

- **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  - New Version
  - Replace Original

^**Note:** When saving as Replace Original document, the current document when found to be containing non-searchable content, will be replaced with a searchable document. The original document will be placed in the location specified in the ‘Backup Location’.

- **Save As – ‘New Version’**

![OCR Backlog - Add Service Wizard](image)

- **Backup** – Location used by contentCrawler safeguard in scenarios whereby a backup of the original source document will be taken before deletion by SharePoint Online if the document version limit has been reached.

  **Note:** contentCrawler will create a backup folder structure in the following order:
  - Backup folder selected or created during service creation process\ file directory reference \Year of file edit date\Month of file edit date\ (e.g. Backup folder \Active\2013\03)
  - It will use the document’s DMS ID as the name of the document and will suffix the name with the version number and minor version number of the document being backed up. (e.g. 43-1.1.tiff)

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the
backup location. Should be specified with the domain name in the format ‘domain name\user name’.

- **Save As – ‘Replace Original’**

![Image of OCR Backend - Add Service Wizard]

- **Backup** – Location used by contentCrawler safeguard in scenarios whereby a backup of the original source document will be taken before being replaced by contentCrawler or deletion by SharePoint Online if the document version limit has been reached.

  **Note:** contentCrawler will create a backup folder structure in the following order:

  - Backup folder selected or created during service creation process\ file directory reference \\Year of file edit date\\Month of file edit date\\
    (e.g. Backup folder \Active\2013\03)
  - It will use the document’s DMS ID as the name of the document and will suffix the name with the version number and minor version number of the document being backed up.
    (e.g. 43-1.1.tiff)

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format ‘domain name\user name’.
10.1.11.8 **Service Settings**

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click *Finish* to complete the creation of the Service.
10.1.12 Worldox GX2 & GX3

To create a new service, go to the File menu - > New and select ‘New’ as shown against the Backlog function or ActiveMonitoring function

The Service Wizard will then prompt you with a series of questions.

10.1.12.1 Process Options

- **Process Types** - Select one or more process type to be performed on documents in the service

  **Note:** It is highly recommended that if licensed for OCR and Compression Modules; that combined processing services are created so ensure documents are processed in the correct order, ie OCR’d and then Compressed.
10.1.12.2 Content Repository Connection Settings

- **Content Repository** - Select the brand of Content Repository (Document Management System) – *Worldox*

- **Worldox User Code** - you should specify a Worldox User code that provides FULL CONTROL permissions to access all documents and Profile groups in your Content Repository.

- **Network Drive(s)** - Select the network drives that are used for Worldox
  - Profile Group location(s)
  - Worldox Admin location
  - Worldox Indexer location

  *Note:* Non-Worldox related network drives will be displayed but do not need to be selected.

- **Get Profile Groups** - Once you have entered the Worldox User Code and selected network drives, click on the ‘Get Profile Groups’ button

- **Profile Group(s)** - Select 1 or more profile groups. Note: Hidden or restricted profile groups of the user will not be displayed

  Click Next to proceed to the next step in the Add Service Wizard.

10.1.12.3 Search Settings

- **Search Dates** – choose the documents you want to assess and process based on the dates they were edited.
Active Monitoring Process

- Lasted Edited on or after – any documents that were edited in any way after the date shown here.

Backlog Process

- Edited on and before – any documents with a last edited date on or before this date
- Edited between – any range of dates

- Document Types – specify the types of documents you would like to search for and assess for processing. You must select at least one type
  - PDF
  - Images - BMP, JPG, PNG, TIFF
  - MSG (Outlook emails)

contentCrawler will check if these emails have any attachments that require processing. It will also check if any of these emails have attachments which are also Outlook emails, and then look in those attachments to see if they are non-searchable.

Select the MSG attachment file types to be included in the search

- PDF
- Images - BMP, JPG, PNG, TIFF

Click Next to proceed to the next step in the Add Service Wizard.
10.1.12.4  Content Repository Search Settings

- **Include versions declared as records** - any document declared as records will also be included in the search. (Available in both Active Monitoring and Backlog)

- **Include read only Documents** – only available in backlog service and means that any read only documents will also be included in the search. (Available in both Active Monitoring and Backlog)

Click Next to proceed to the next step in the Add Service Wizard.
Assess Settings

OCR Assessment (Only available when Process Type OCR select for service)

OCR PDF documents with average characters per page less than or equal to [x] - In some cases you may have PDF documents (either profiled separately or as attachments to emails) that have non-searchable content but also have text. For example, you may have a scanned page from a photocopier which has had a page number applied to it. In this example the document does have text (the page number) but certainly the rest of the document is not text-searchable. To cover these types of documents, you can specify the minimum average number of characters per page which must be found by the contentCrawler assessing tool for the document to be classified as text-searchable. The default is an average of 40 characters per page. The assessing tool will analyse the document to calculate the average characters per page, this is done by obtaining the total number of characters within the whole document and divide this by the number of pages. If the average character count per page exceeds 40, the document will be classified as not requiring OCRing. Example if the document has 10 pages, and you set this limit as 40 characters, that means there must be more than 400 characters in the whole document to be assessed as not requiring OCR’ing.

Compression Assessment (Only available when Process Type Compress select for service)

Set the minimum file size of documents which are to be processed for compression, the default is 500KB and only compress documents which contain image content which is greater than 50% of the overall file size.

Note: Compression can only be gained on documents which contain image content as the compression process will resize and/or resample images contained within the PDF.
The administrator can opt to ONLY process documents which are marked for OCR (default) or can select “Include documents that do not require OCR”, which will also assess ALL documents including existing text searchable PDF files for compression.

**Note:** This option is only available when used in conjunction with the process type OCR. A compression only services will assess ALL IMAGE, PDF and MSG regardless of text search ability

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before any processing is run on those documents that are assessed as needing OCRing and/or Compression. It is recommended that you select this when first installing and testing contentCrawler so that you are comfortable that the software is working as required.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.12.6  Process Settings

Allows the administrator to use the default compression setting (Recommended), which are a balance of small size and High quality or use the slider to change the compression ratio.

The Advance option allows the administrator greater control over the compression settings. Allowing different quality and compression types to be set. Refer to Compression Advance Settings.

Note: Compression settings are set per service and each service can be configured with unique compression settings if required.
The administrator also has the ability to control if compressed documents need to meet a compression gained threshold before being saved back to the document repository.

Administrators may only require documents which have gained 15% file size reduction to be saved back to the DMS. This threshold can be configured using the 'Save compressed documents reduced by at least \[x\] % of file size. By default this is set to 15%. Please refer to the Administrator Guide for full details on compressed document save handling.

- **Hold for Review** – With this option selected, the documents will be held for review by the contentCrawler Administrator, before the result document is re-profiled back into your Content Repository. This again allows the Administrator to review the original source document, the result document after OCRing and/or Compression and compare the output, prior to allowing contentCrawler to proceed to save the document.

**Note:** If the contentCrawler license key indicates you have a trial mode of contentCrawler, a maximum of 100 documents assessed as needing to be processed (OCR and/or compressed) can be pass through the process stage and saved back to the Content Repository.

Click Next to proceed to the next step in the Add Service Wizard.
10.1.12.7 Content Repository Save Settings

- **Save As** – You must select how you want to re-save the document after the OCRing has been completed.
  - Worldox – you may choose;
    - New Version
    - Replace Original

  ^Note: When saving as Replace Original document, the current document, when found to be containing non-searchable content, will be replaced with a searchable document. The original document will be placed in the location specified in the 'Backup Location'.

- **Save As – ‘New Version’**
• **Save As – ‘Replace Original’**

![Save As ‘Replace Original’](image)

- **Backup Location** – When a document is found by contentCrawler to be non-text searchable, the document will be replaced with the OCR’d version of the document.

  **Note:** contentCrawler will create a backup folder structure in the following order:

  - Backup folder selected or created during service creation process \Profile Group Name\Year of backup date(YYYY)\Month of backup date(MM)\DocumentID\Version Number\ (e.g. Backup folder \Active\2013\04\000012345\1\)
  - It will use the document’s DMS ID as the name of the document and will suffix the name with _CCID#####.ext. Where ##### is a unique number (e.g. 000012345_CCID2456.tiff)

- **Credentials (User name & Password)** - you should specify a full administrator user name and password that provides FULL CONTROL permissions to access the backup location. Should be specified with the domain name in the format ‘domain name\user name’.

  Click Next to proceed to the next step in the Add Service Wizard.
10.1.12.8 Service Settings

- **Name** – Specify the name for this Service. You can provide any name you wish.

- **Start service now** – Select this checkbox if you wish to start the service immediately – you can later switch on this service from the Dashboard.

- **Trial Mode (only available for Backlog)** – If the service contains a Trial Module, or if you wish to only ‘audit’ your documents but not process them, you may select the ‘Trial Mode – Enabled’ option. Once selected you can specify the amount of time that contentCrawler will spend assessing your documents. This assessment is completed by randomly sampling documents in your DMS so that not every single document has to be tested. It is recommended that if you have more than 100,000 documents approximately, leave the percentage of documents at 10% and you will quickly get a clear percentage assessment of how many documents require processing. Alternatively you can specify the percentage of documents or how long you want the testing to occur for.

Click *Finish* to complete the creation of the Service.
## 10.2 Compression Advance Settings

### 10.2.1 Colour Images

<table>
<thead>
<tr>
<th>Setting</th>
<th>Default</th>
<th>Setting Options</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>DownSample</td>
<td>Bicubic Downsample</td>
<td>• Average Downsampling&lt;br&gt;• Subsampling&lt;br&gt;• Bicubic Downsample</td>
<td>• Average Downsampling will average the pixels in an area and replace the whole area with an average pixel color.&lt;br&gt;• Subsampling will select a pixel in the center of a sample area and replace the whole area with the selected pixel color. Faster conversion times are achieved with this method.&lt;br&gt;• Bicubic Downsample is a weighted average used to determine the pixel color. More often than not this will return favorable results but is slower in its conversion methods.</td>
</tr>
<tr>
<td>DPI Settings</td>
<td>150 dpi for image above 225 dpi</td>
<td>• Configurable range</td>
<td>• Reduces images above 225 dpi to 150 dpi&lt;br&gt;• Images below 225 dpi will not be reduced</td>
</tr>
<tr>
<td>Compression</td>
<td>JPEG2000</td>
<td>• Retain Existing&lt;br&gt;• JPEG2000&lt;br&gt;• JPEG&lt;br&gt;• ZIP</td>
<td>• Retain Existing&lt;br&gt;• JPEG2000&lt;br&gt;• JPEG is suitable for Color or Grayscale images and uses lossy compression methods, meaning it removes image data which may impact the quality of the image.&lt;br&gt;• ZIP is the preferred method when dealing with large areas of single colors or black and white images. This is a lossless method.</td>
</tr>
<tr>
<td>Scale</td>
<td>3</td>
<td>• 1 – Smallest Size&lt;br&gt;• 2&lt;br&gt;• 3&lt;br&gt;• 4&lt;br&gt;• 5 – Highest Quality</td>
<td>• Sets the balance point between file size and quality</td>
</tr>
<tr>
<td>Convert to Grayscale</td>
<td>Off</td>
<td>• On&lt;br&gt;• Off</td>
<td>• Converts color images to Greyscale</td>
</tr>
<tr>
<td>Convert to Monochrome</td>
<td>Off</td>
<td>• On&lt;br&gt;• Off</td>
<td>• Converts images to black and white</td>
</tr>
</tbody>
</table>
### 10.2.2 Greyscale Images

<table>
<thead>
<tr>
<th>Setting</th>
<th>Default</th>
<th>Setting Options</th>
<th>Info</th>
</tr>
</thead>
</table>
| DownSample      | Bicubic Downsampling | • Average Downsampling  
• Subsampling  
• Bicubic Downsampling | • Average Downsampling will average the pixels in an area and replace the whole area with an average pixel color.  
• Subsampling will select a pixel in the center of a sample area and replace the whole area with the selected pixel color. Faster conversion times are achieved with this method.  
• Bicubic Downsampling is a weighted average used to determine the pixel color. More often than not this will return favorable results but is slower in its conversion methods. |
| DPI Settings    | 150 dpi for image above 225 dpi | • Configurable range                                                                 | • Reduces images above 225 dpi to 150 dpi  
• Images below 225 dpi will not be reduced |
| Compression     | JPEG2000          | • Retain Existing  
• JPEG2000  
• JPEG  
• ZIP | • Retain Existing  
• JPEG2000  
• JPEG is suitable for Color or Grayscale images and uses lossy compression methods, meaning it removes image data which may impact the quality of the image.  
• ZIP is the preferred method when dealing with large areas of single colors or black and white images. This is a lossless method. |
| Scale           | 3                 | • 1 – Smallest Size  
• 2  
• 3  
• 4  
• 5 – Highest Quality |                                                                                           |
| Convert to Grayscale | Off               | • On  
• Off | • Converts color images to Greyscale |
### 10.2.3 Monochrome Images

<table>
<thead>
<tr>
<th>Setting</th>
<th>Default</th>
<th>Setting Options</th>
<th>Info</th>
</tr>
</thead>
</table>
| DownSample    | Bicubic Downsample           | • Average Downsampling  
• Subsampling  
• Bicubic Downsample | • Average Downsampling will average the pixels in an area and replace the whole area with an average pixel color.  
• Subsampling will select a pixel in the center of a sample area and replace the whole area with the selected pixel color. Faster conversion times are achieved with this method.  
• Bicubic Downsample is a weighted average used to determine the pixel color. More often than not this will return favorable results but is slower in its conversion methods. |
| DPI Settings  | 300 dpi for image above 450 dpi | Configurable range                                   | • Reduces images above 225 dpi to 150 dpi  
• Images below 225 dpi will not be reduced |
| Compression   | • JBIG2                      | • Retain Existing  
• JBIG2  
• CCITT Group 3  
• CCITT Group 4  
• LZW  
• Run Length | • Retain Existing  
• JBIG2 in general will have more favorable results on images over CCITT.  
• CCITT is only available for black and white images. Group 4 will provide a slight advantage over Group 3.  
• LZW uses a table-based algorithm commonly used for GIF and TIFF file formats  
• Run Length (RLE) uses a lossless compression method, and returns favourable results for images with large areas of solid black or white. |
| Scale         | 3                            | • 1 – Smallest Size  
• 2  
• 3  
• 4  
• 5 – Highest Quality | • Sets the balance point between file size and quality |
10.3 Processing steps in Backlog or ActiveMonitoring

When a service operates, there are three basic stages in the processing of documents.

10.3.1 Stage 1. Search & Assess

In this stage, contentCrawler will firstly search for the latest version or revision depending on DMS, of all documents that meet the criteria defined when the service was created. The document search criterion is shown on the screen. The search criteria include the document type and date last edited.

Note. Open Text eDocs DM, sub-versions and attachments will be ignored. Worldox, sub-versions will be ignored.

This returns a quantity of documents which is shown in the example screen as “100 documents assessed”

The quantity in this example of 100 is not the total number of documents in your Content Repository - it is those documents matching the search criteria. So you may have 10,000 documents in your Content Repository, but 100 match the search criteria.

Once these 100 documents are found as a list of documents, contentCrawler will firstly confirm if it has information on this document in its own database (storing only enough information to uniquely identify the document including date last edited). If this document was previously assessed and either confirmed as being assessed complete (no processing required), or it was previously assessed processed and saved, then the document is immediately dismissed as requiring any assessing. Of course if the document was modified in your Content Repository since the last time contentCrawler processed it, the document will be fully assessed again now.

Should the document require assessing, contentCrawler will assess the document based on the assessment criteria for the processing types selected in the service.

The ‘Exceptions’ indicate provides information on which documents could not be assessed for various reasons. This could include reasons such as:

- The document is corrupt – it may be profiled as a TIFF image but may in fact be a Word document
- The Content Repository cannot retrieve the document – indicates a problem with your Content Repository database. For example Authentication problems or communication failures.
- The PDF document may have a password preventing it being viewed or modified
- For File System content repository; the original document or folder it is contained within, was deleted, renamed or moved
If you are in 'Trial Mode' when you define the Process, only a random sampling of documents in the 100 documents to be assessed will actually be assessed. Importantly, this sampling is completely random, so you can be sure that even if only 10% of your documents are being assessed, the statistics provided to you regarding the quantity of documents that are currently not searchable will be highly accurate. This is a way of quickly and reliably giving you a good understanding of the size of the problem you are facing – the number of documents in your Content Repository that are not text searchable.

If you have selected the ‘Hold for Review’ checkbox, no documents will be passed onto the process stage until the documents are either individually released or the ‘Hold’ option is unchecked.

You will find that the Search and Assess stage is reasonably fast. As an example, if you have 1,000,000 documents in your Content Repository and 10% of these documents are PDFs, images, or emails, (100,000 documents), contentCrawler will take about 30 minutes to create the list of 100,000 documents, and then around ¼ second per document to assess each one and determine whether it needs to be processed.

If you are using the 'Trial Mode' to assess the amount of unsearchable content, you will be provided some overall statistics at the end of the Assessment, which can also be saved to a comma delimited file to review in Excel.

In this example, of the 100 documents in the search criteria, only 10% of them are to be assessed. Of that subset, 100% of the documents are not searchable. This means that 100 out of the total of 100 most likely will require processing.

The Search & Assess process goes through a number of stages that can be reported on. You should be aware that the 'Report' tab will only display the current status in the processing of the document. The statuses of processing a document in the Search & Assess stage are:

- **Queue** – document has been found by the Search service and is queued to be assessed.
- **In Progress** – document is currently being assessed to determine if the document is already text-searchable and/or requires compression.
- **Completed** – document was assessed and found to already be text-searchable and/or does not require compression
- **Held** – document has been assessed as requiring processing but has been held at this stage and not passed to the Process stage.
- **Exceptions** – document failed the assessment stage – the document may be unreadable, corrupted, etc.
10.3.2 Stage 2: Process

The Process stage will accept any documents flagged in the Search and Assess stage which require processing.

This will process documents based on the process type configured during the service creation.

If the ‘Hold for Review’ checkbox has been selected, these documents will be stored locally on your computer so that the results of the process can be individually checked before the document is re-profiled. These documents are stored in the Cache. Refer to section 6.1.5 Configure Document Cache for more information on how to set a location for the cache. It is important to ensure the cache directory has enough storage space to hold large amount of data when opting to use the Hold for Review option.

Exceptions are noted where there has been a problem OCRing and/or compression the document. This may be because the document cannot be correctly retrieved from the Content Repository, due to security restrictions on the document prohibiting content access or internal corruption in the documents itself.

The Process stage can take some time to run, depending on the speed of your computer and the size of the documents.

The Process stage goes through a number of stages that can be reported on. You should be aware that the ‘Report’ tab will only display the current status in the processing of the document. The statuses of processing a document in the Process stage are:

- **Queue** – document has been found by the Search process and is queued to be processed (OCR’d and/or Compressed).

- **In Progress** – document is currently being processed (OCR’d and/or Compressed), note In Progress % are only shown on the Live View

- **Held** – document has been processed successfully but has been held at this stage and not passed to the Save stage.

- **Exceptions** – document failed the process stage – the document may be unreadable, corrupted, in use, etc.
10.3.3 Stage 3. Save

This stage will re-profile the documents based on the criteria you defined when you created the Service.

This will also indicate which documents failed to save in the ‘exception’ area.

Documents may fail the save stage mainly if the document is currently in use by another user, or has already changed since in the document was processed (e.g.: another user on the network changed the document at the same time it was being processed by contentCrawler).

The saving process is generally quite quick – typically around 1 second per document.

The Save process goes through a number of stages that can be reported on. You should be aware that the ‘Report’ tab will only display the current status in the processing of the document. The status of processing a document in the Save stage are:

- **Queue** – document has been finished by the process and is queued to be saved.
- **In Progress** – document is currently being saved.
- **Completed** – document has now been successfully saved to the Content Repository.
- **Exceptions** – document failed the Save stage – the document may have been in use or some other error occurred.

10.4 Monitor and Review a service

Once a process has commenced, you may review its progress at any time through the Dashboard.

Simply go to the Dashboard menu item and then click on the Hyperlink for the Backlog process that you wish to review.
10.5 Service Detailed Dashboard

The Dashboard details screen show the current and past activity for this particular contentCrawler Service.

The many features of this dashboard are explained below so that you can monitor visually the overall process of the service as well as check on specific documents and the results of its processing.

10.5.1 Detailed Dashboard Status

This section of Dashboard gives you an overall summary of progress.

10.5.1.1 **On/Off Switch**

This can be used to switch the Service on and off. Switching the service off stops all processing defined in the service. The example shown here shows the service is currently off.

10.5.1.2 **Status Bar**

Shows in a visual sense what percentage of the total task defined for the service has been achieved.

10.5.1.3 **Status Description**

This shows how many documents have been assessed (50 out of a total of 520 documents that match the search criteria) and the assessment stage has been completed. By clicking on the complete hyperlink you can view a report on the report tab showing those documents that were assessed and require no processing or have been saved successfully.

This status description also shows any documents that are being held for review (which can be viewed by using the review hyperlink), and any documents that have been processed and saved.
10.5.2 Detailed Dashboard Setup Parameters

This section of the screen shows the settings for this particular Service,

10.5.2.1 Search
Indicates what documents are to be searched and assessed including the document file type and last edited date

10.5.2.2 Operation
Defines what process is to be performed and how the document is to be re-profiled into your Content Repository once the document has been processed. This will describe how the document will be saved back into the content repository.
10.5.3 Detailed Dashboard - Search & Assess

This section of the screen provides an up to the minute progress of the Search and Assess stage. It indicates the total number of documents that are determined that require Assessing. As the documents are assessed, the total number of documents, including the total number of pages which have been assessed will increment.

The total number of documents which require processing will be displayed.

A breakdown of each of the processing types will be displayed, showing the total number and total pages of the documents which require OCR and/or Compression. This information is also included in the Audit report information.

By clicking on the exception(s), or Search & Assess or Review hyperlinks, you will be taken directly to the Report tab and the report parameters automatically selected for the documents described by the hyperlink.

Please refer to section 8.5.1 Search & Assess of this manual for detailed information about how the Search & Assess operates.

The Search & Assess hyperlink will display those documents that have been found as part of the Search phase and are being or have been assessed.

Exceptions status indicates those documents that had an error when the document was attempted to be assessed for non-searchable content.

The Review status indicates that the document has been ‘Held’ and can be reviewed by you so that you can decide if you wish to allow it to continue to the next stage of the processing.

10.5.4 Detailed Dashboard – Process

The Process stage will Process any document that has successfully passed the assessment criteria set for the process types.

This section will display the total number and total file size of documents which have been processed.

A breakdown of each of the processing types will be displayed, showing the total number and total pages of the documents which have been processed for OCR and/or Compression.
All values will increment the totals as each document is processed.

The Process, exceptions and Review hyperlinks will take you directly to the Report tab and the report parameters automatically selected for the documents indicated by the hyperlink.

The PROCESS hyperlink shows those documents that are to be, or have been processed

Exceptions status indicates those documents that had an error when the document was attempted to be OCR’d and/or compressed. The Review status indicates that the document has been ‘Held’ and can be reviewed by you so that you can decide if you wish to allow it to continue to the next stage of the processing.

10.5.5 Detailed Dashboard – Save

This stage will save the documents that have been processed based on the configuration for this service.

This section will display the total number of documents which are to be Saved and will increment the total as each document is Saved. For Service which include Compression processing, the total file size saved will also be displayed.

The Save hyperlink shows you in a report those documents that are to be, or have been saved. This would only be documents that have been successfully processed during the process stage.

The exceptions hyperlink will take you directly to the Report tab, displaying those documents that failed during the saving process
10.5.6 Live View Reporting Pane

The Live View reporting pane provides literally a live view of any documents currently being dealt with by contentCrawler. It provides real-time reporting on the last 25 actions performed by contentCrawler. This will describe the document path, number, version number, document type, and the Status of the current process. That status could be In Progress, Saving, Completed, Exception, etc.

If you want to report on a larger range of documents and track how a document progressed through contentCrawler, or all documents that have been processed, then you should use the ‘Report’ tab.

When you look at the Live View tab, it shows at the top of the screen the documents currently being worked on. Each time the screen refreshes, the ‘Status’ will change to show you progress. During the Process stage the status column will show an **In Progress %** which will increment in accordance with process being performed on the document and will increment each time the screen refreshes.

It is important to note that concurrent document processing are shared across ALL Active Services, the number of documents showing in progress at the ‘Process’ stage for the particular service being viewed may fluctuate, refer to section 4.4.6 Multi Document Processing. The progress percentage is **NOT** displayed in the Report Pane.

Once the next document is being processed, the current document will move down the live view. You may only view the most recent 25 documents being processed in the Live View.

You cannot sort the columns in the Live View tab as there is never more than 25 lines, but you can change the width of the columns by selecting the column divider and dragging left or right.

There is no ability to right-click on a document in the Live View tab to view the source document or results. You may only do this when viewing the document in the ‘Report’ pane.

The Live pane will display the following columns:

**Document ID** - Original document name or ID, depending on the DMS

**Type** - Original document file type - BMP, JPD, MSG, PDF, PNG, TIFF and DXL if supported

**Pages** - Original document page count. MSG and DXL page counts will be calculated on the File type attachments matching the search criteria.
Original Size - Original document total file size on disk. MSG and DXL file size is the overall file size not the attachment size

^contentCrawler may be unable to determine the Pages or Size on Exception documents

Result Size - Result document total file size on disk. MSG and DXL file size is the overall file size not the attachment size (Only available for Compression service)

OCR – Status flags used to indicate if a document does or does not require processing for OCR (Only available for OCR service)

Compress - Status flags used to indicate if a document does or does not require processing for compression (Only available for Compression service)

<table>
<thead>
<tr>
<th></th>
<th>The document has not yet been assessed for OCR or compression.</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>The document has been assessed and does not require OCR/compression.</td>
</tr>
<tr>
<td>!</td>
<td>The document has been assessed as requiring OCR/compression, and is waiting to be processed.</td>
</tr>
<tr>
<td>✓</td>
<td>The document has been assessed as requiring OCRd/compressed, and has successfully OCRd/compressed.</td>
</tr>
<tr>
<td>✓</td>
<td>The document has been assessed and requires compression, but has failed to reduce by the percentage specified in the service creation wizard.</td>
</tr>
<tr>
<td>✓</td>
<td>The document has been assessed and requires OCR/compression, but has failed to OCR/compress due to an exception.</td>
</tr>
</tbody>
</table>

Stage – Current stage of the document – Assess, Process and Save

Status – Status at current stage; Queued, In Progress (%)*, Held, Complete and Exception

* A progress percentage is only displayed during the OCR Stage, Assess and Save will only show 'In Progress’

Last Updates – Last Date/Time the document was processed by contentCrawler
10.5.7 Report Pane

The Report pane allows you to view all documents that have been processed by contentCrawler, and see its current status. A document will only appear once in the Report view showing its most recent status and stage. (i.e.: it will not show a separate entry for each stage in the processing of the document).

Typically you will be shown the Report pane when you click on one of the Hyperlinks in the Search, OCR and Save panel. In this case, the Report pane will automatically select those documents required to be viewed based on the hyperlink you chose. Alternatively, you can select the documents you wish to search for using the Report Filter tab at the top left of the screen.

At any time, you can select the Report pane by clicking on the Report tab, regardless of whether the Service is actually active at the time or not.

The Report pane will display the following columns

**Document ID** - Original document name or ID, depending on the DMS

**Type** – Original document file type - BMP, JPD, MSG, PDF, PNG, TIFF and DXL if supported

**Pages** – Original document page count. MSG page count will be calculated on the File type attachments matching the search criteria

**Original Size** – Original document total file size on disk. MSG and DXL file size is the overall file size not the attachment size

```
^contentCrawler may be unable to determine the Pages or Size on Exception documents
```

**Result Size** - Result document total file size on disk. MSG and DXL file size is the overall file size not the attachment size (Only available for Compression service)

**OCR** – Status flags used to indicate if a document does or does not require processing for OCR (Only available for OCR service)

**Compress** - Status flags used to indicate if a document does or does not require processing for compression (Only available for Compression service)

<table>
<thead>
<tr>
<th>Status</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>The document has not yet been assessed for OCR or compression.</td>
</tr>
<tr>
<td>N/A</td>
<td>The document has been assessed and does not require OCR/compression.</td>
</tr>
<tr>
<td>✔️</td>
<td>The document has been assessed as requiring OCR/compression, and is waiting to be processed.</td>
</tr>
<tr>
<td>✔️</td>
<td>The document has been assessed as requiring OCRd/compressed, and has successfully OCRd/compressed.</td>
</tr>
</tbody>
</table>
Stage – Current stage of the document – Assess, Process and Save

Status – Status at current stage; Queued, In Progress*, Held, Complete and Exception
* In Progress at the Progress stage will not show a progress percentage this is only available in the Live Pane

Exception Error – Exception error occurred during processing. Note this column is only available in the Report pane and is only display if the Report Filter includes Exception documents.

Last Updates – Last Date/Time the document was processed by contentCrawler

At the bottom of the Report pane, there are buttons to allow go to the First screen, previous screen, or to view the many separate pages, each showing a quantity of documents.

You can also control how many documents you can view at any one time by using the ‘View’ tab to set how many documents to view at the one time. If there are more documents to show that can fit on one screen, you will be able to refresh.

The Refresh button will draw the current window to check for any documents that have changed their stage of processing.

The Report Filter tab allows you to search for a range of documents. When using a document management system, you may also search for one document specifically if you know the Document number.

If you wish to search for only one Document, just type in the Document unique identified in the ‘Search Doc #’ field and press the ‘Enter’ key on your keyboard or click on the magnifying glass icon.

You cannot sort the columns in the Report View tab, but you can change the width of the columns by selecting the column divider and dragging left or right. Alternatively you can export the report information to a .csv file using the Report Export option. Using the Report Filter options to return the desired data, select to export All pages or use the Pages setting to export certain pages or page ranges.
The Release Documents button is only able to be selected when you choose to view documents with a status of Held (using the Report Filter or appropriate ‘Held’ hyperlink). In this case, clicking on the ‘Release documents’ button will release all held documents. Please note that you can choose to individually release a document by right-clicking on the document or multiple documents in the Report view and selecting the option to ‘Release Held document’.

When you right click on a document in the Report pane, you can then choose to view the actual source document in whatever application it was created in, view the result of the process as long as you have a version of Acrobat installed or deal with an error. This allows you to review the document before and after it has been processed.

If the document is held or queued for processing, but you do not want it processed, you can select the ‘Remove Document from Service’ option.

If the document you select shows as status ‘Exception’, the error that occurred whilst processing that document will be show in the ‘Exception Error’ column in the report pane, this information can also be included in the report export.

Note this column is only available in the Report pane and is only display if the Report Filter includes Exception documents.

Alternatively you can right click on that document to view the error, and re-submit it for processing. For example, the document may have failed processing as it was locked by another user. Once the document has been check-in, the document can be specifically re-attempted.

A Re-Attempt All button is also available and can be selected when you choose to view documents with a status of Exception (using the Report Filter or appropriate ‘Exception’ hyperlink). In this case, clicking on the ‘Re-Attempt All’ button will re-attempt all exception documents for reprocessing. Please note that you can choose to individually re-attempt a document by right-clicking on the document or multiple documents in the Report view and selecting the option to ‘Re-Attempt Document’.
The ‘Remove Document from Service’ option is also available on documents with ‘Exception’ status, allowing known problem documents to be removed from the current service and all other services, including any new services created. Important: Once removed contentCrawler will not find the document again for processing.
11 Upgrading contentCrawler

Upgrading contentCrawler is a very straightforward process. It retains all your previous settings when you upgrade, including license keys and processing history.

If using Open Text eDocs DM – Save as New Version, to OCR Image files, and are upgrading from 1.7 or any prior version, please contact DocsCorp Support (support@docscorp.com) for advice on how to upgrade.

11.1.1 Backup Internal Database

It is a good idea to backup your existing database so records of previously processed documents are retained. If you lose that information, then contentCrawler will need to analyse all your documents again to determine if they are text-searchable - a time consuming process.

The contentCrawler upgrade process will automatically perform a backup of the contentCrawler database (ccdb.sdf), however will not backup the working Cache files references in step 3.

**Step 1** – Quit out of the contentCrawler Administration Console

**Step 2** – Terminate the “DocsCorp contentCrawler Server” - to do this, select Start Menu - > Control Panels – Administrative Tools – Services

Select the 'DocsCorp contentCrawler Server' and choose the 'Stop the service' option.

**Step 3** – Make a backup of the contentCrawler data directory. If something goes wrong, you can restore this directory. You will find the Data directory where the Cache is stored. Read the section on Cache settings for further information, refer to section 6.1.5 Configure Document Cache in this manual. Ideally you should back this up to another directory or drive. The default location for this folder will be found at C:\Program Files\DocsCorp\contentCrawler\Data.
11.1.2 Perform the Upgrade

Follow these steps to ensure contentCrawler upgrade completes successfully

1. Download the new contentCrawler MSI installer to a location on your contentCrawler environment

2. Open the contentCrawler administrator console

3. Pause all current Services by switching the Service from On to Off

4. Ensure all documents showing ‘In Progress’ complete their current processing stage

5. Once all services have been paused and no documents are currently ‘In Progress’, close the contentCrawler Administration console

   **Note:** Background processes may take additional time to complete their tasks, it would be advisable to double check these tasks have been completed via Task Manager before upgrading. Ensure there are no instances of Recognition.exe or FREngine.exe still running, wait until these have finished before performing the upgrade.

6. Carry out the upgrade by double-clicking on the contentCrawler .msi.

   This will automatically uninstall any older version of contentCrawler, install the new version, and upgrade any data files required. At completion of the installation, the contentCrawler windows service will be restarted

7. Once complete, launch the contentCrawler Administrator console.

   It is highly recommended that once the installation has completed, you check the processes that you have configured with contentCrawler to ensure that any new options provided in this newer version are correctly configured.

8. Restart all Services by switching the Service from Off to On

   **Note.** If you have installed to custom location, such as D:\Apps, this will be maintained and you will not be required to specify this again during the installation Wizard. If you require contentCrawler to be installed to a different location when a manual uninstall via add/remove programs will be required before running the msi to perform a new install.

   Please also be aware that the contentCrawler database is located in the installation directory, if you require your existing service information and data to be maintained. It is recommended that the 'Data' directory from your original contentCrawler installation is copied to the new location before the manual installation is performed.
11.1.3 Moving contentCrawler Installation

In some cases, you may wish to move your installation of contentCrawler from one computer to another.

Follow these steps to ensure an easy transition

9. Install contentCrawler on the new computer – no need to enter any License Keys.

10. Stop the DocsCorp contentCrawler Service on both the new computer and the original installation computer

11. Copy the ‘Data’ directory in C:\Program Files (x86)\DocsCorp\contentCrawler from your original contentCrawler installation and replace the ‘Data’ directory on this new computer in C:\Program Files (x86)\DocsCorp\contentCrawler.

12. Start the DocsCorp contentCrawler service on the new computer

13. Remove the contentCrawler from the original machine from add/remove programs
12 Best Practise

12.1 Perform a contentCrawler Server Reboot

Administrators may from time to time be required to run routine reboots on the server where contentCrawler is housed. In order to do so, it is advisable to run through the following steps prior to the reboot to ensure that all contentCrawler services resume successfully on restart.

1. Open the contentCrawler administrator console
2. Pause all current Services by switching the Service from On to Off
3. Ensure all documents showing 'In Progress' complete their current processing stage
4. Once all services have been paused and no documents are currently 'In Progress', close the contentCrawler Administration console

**Note:** Background processes may take additional time to complete their tasks, it would be advisable to double check these task have been completed via Task Manager before initiating the reboot. Ensure there are no instances of Recognition.exe or FREngine.exe still running, wait until these have finished before performing the reboot.

5. Carry out the server reboot
6. On restart, launch the contentCrawler Administrator console
7. Restart all Services by switching the Service from Off to On
13 Worldox – Changing the windows service user password

On a Worldox environment, contentCrawler requires the DocsCorp contentCrawler Server windows service to run under the user context to allow access to all mapped drives to the Worldox profile groups have been configured, this will generally be a user with full administrator rights. Clients may impose security measures which requires new passwords to be set periodically for admin accounts. Below are details on how to change the password successfully to ensure contentCrawler continues to run without access issues.

Follow these steps to ensure an easy transition

1. Launch the contentCrawler Administrator Console
2. Pause all running Services
3. Ensure all “In Progress’ jobs complete before proceeding
4. Close the contentCrawler Admin Console
5. Navigate to windows services
6. Stop the DocsCorp contentCrawler Server Service on the new computer
7. Once stopped, right click on the service and select properties
8. Under Log On, change the password

![Password Change](image)

9. Apply changes and Ok
10. Start the DocsCorp contentCrawler Server Service on the new computer
11. Remove the contentCrawler from the original machine from add/remove programs
12. Launch the contentCrawler Administrator Console
13. Start all Services