



- Increase organizational productivity
- Simplify management of image-based documents
- Reduce non-compliance risks
- Increase efficiency through automation
- Leverage investment in DMS and search technology
- Reduce costs managing OCR technology
- Report on processing via console or SQL Server database
- Supports recognition of 180+ languages

contentCrawler is an integrated bulk processing framework that assesses documents for OCR and compression services.

contentCrawler assesses and analyzes documents in a content repository based on search criteria, as well as text and compression thresholds configured by IT Administrators. Documents are processed based on the service type (OCR, Compression or both), and saved back into the content repository. This is an automated server-based process that does not impact the desktop user.

### MAKE EVERY DOCUMENT SEARCHABLE

Businesses have invested heavily in document management systems as well as in search technology to ensure they have instant access to the documents they need when they need them.

Despite this investment, 30% of documents in content repositories may be non-searchable and, therefore, invisible to search technology.

Failure to locate a business-critical document can undermine efficiency and productivity as well as put an organization's reputation and financial well-being at risk.

The contentCrawler framework can identify non-searchable content in a document management system database or a subset of documents based on specific queries.

The OCR module converts this content to text-searchable PDFs, saving them back into the content repository as new or replacement documents.

*"We use contentCrawler to ensure that newly profiled and legacy PDFs are fully text-searchable. DocsCorp has worked closely with us and has been very responsive to our requests for program enhancements."*

**Jeff Hutchinson:**  
**Mendes & Mount,**  
**LLP - Director**  
**of Information**  
**Technology**

*"contentCrawler has uncovered a range of documents, including PDFs that had previously not been searchable within our DMS. The solution has greatly enhanced our ability to find documents quickly with the use of our DMS search functionality."*

**Mark Turner: Lubbock Fine -**  
**Managing Partner**

### REDUCE FILE SIZE

Storage space in a document management system can be expensive. Large files can be costly to download and slow to send by email.

The contentCrawler Compression module can help reduce storage costs, speed up file transfers when downloading or sending via email by reducing the size of the files in your document management system.

The contentCrawler Compression module enables Administrators to compress image and PDF documents. Converting image documents to PDF and applying compression and downsampling to the files reduces overall file size.

### MULTI-PROCESS SERVICES

IT Administrators are able to combine the OCR and Compression modules into a single service.

Image documents will be converted to text-searchable PDFs to ensure the highest image quality. The Compression module then reduces document file size through compression and downsampling.

### EFFICIENCY THROUGH AUTOMATION

contentCrawler is an end-to-end automated solution that runs 24/7 without staff intervention. Staff do not have to worry about OCR or Compression processes or workflows. Instead, contentCrawler works in Backlog mode for legacy documents and Active Monitoring for recently-profiled documents.

It can work in both modes simultaneously.

<b>MULTI-FUNCTION PROCESSING</b>	OCR documents to produce text-searchable PDFs (OCR module) Compress PDF documents to reduce file size (Compression module)	<b>SYSTEM REQUIREMENTS</b>  <b>Operating Systems</b> Microsoft® Windows Server® 2019, 2016, 2012 R2, 2012*, or 2008 R2 SP1* - all in 64-bit environment only * Not supported on Server Core Role MS .NET Framework 4.5/4.5.1  <b>Hardware</b> 8 GB RAM 100 GB free disk space 1-2 GB per CPU core over 4* * Recommended: 4 dedicated CPUs  <b>INTEGRATIONS</b> File System HP TRIM/Records Manager iManage Work Micro Focus Content Manager MS SharePoint MS SharePoint Online (O365) NetDocuments OpenText Content Server OpenText eDOCS ProLaw Worldox  The application makes use of the following recognition technologies: ABBYY® FineReader® Engine 11.0 © 2008. FINEREADER, ABBYY & ABBYY FineReader are registered trademarks of ABBYY Software Ltd.  'contentCrawler' and the contentCrawler logo are trademarks of DocsCorp Group Pty Ltd.  contentCrawler technology is protected under US Patent 8745084
<b>FAST PROCESSING</b>	Concurrent processing utilizes available CPU cores Default 4 CPU cores, additional licensing for 8, 16, and 32 CPU cores Simultaneous Search, Assess, Process, and Save stages	
<b>CONTENT REPOSITORY SEARCH</b>	Define searches to identify image documents, PDFs, emails with attachments Supports TIF, JPG, PNG, and BMP image types Refined searching on date ranges for Active Monitoring and Backlog modes Supports multiple content repository databases or libraries	
<b>MONITORING MODES</b>	Automates workflows to make documents searchable and/or compressed Assesses and processes newly-profiled or edited document profiles on a regular schedule using Active Monitoring mode Legacy document handling and processing using Backlog mode	
<b>ASSESSES DOCUMENTS</b>	Assesses content in a repository for text searchability (OCR module) and/or compression capability (Compression module)	
<b>OCR MODULE</b>	Documents are OCRd to generate PDFs with a hidden text layer Intelligent OCR technology ensures document fidelity No requirement for a text file separate to the image or PDF file; hidden text layer is searchable and indexable Use Search feature in PDF viewer to find and review exact content Multi-language recognition of over 180 languages (including Arabic, Hebrew, Thai, Vietnamese, and Yiddish); unlimited page processing Select up to 16 languages for OCR recognition with no impact on processing speed	
<b>COMPRESSION MODULE</b>	Compresses using standard JPEG, JPEG2000, and JBIG2 formats Resizes and downsamples PDF image documents to reduce file size Auto-converts image files to PDF prior to compression, and converts to text-searchable PDF when used with OCR module Configurable compression and file size reduction settings Reduces risk of download and email size limitations	
<b>SAVE TO DMS</b>	Uses DMS API/Webservices calls when saving documents back, honoring permissions and access rights Image documents rendered to PDFs and compressed before Save Replace email attachments with processed and compressed PDFs before Save Save options can include: as a New Version, Related Document, New Rendition or as an Attachment depending on your document management system Add a Prefix or a Suffix to a new version document description in iManage	
<b>AUDIT AND REPORTING</b>	Centralized Administration Console for monitoring, configuring, and reporting Hold for Review options prior to Processing and Save stages Email notifications for periodic processing statistics and error reporting Exception reporting and 'Not supported' list filtering Optionally use a Microsoft SQL Server database for processing and reporting	
<b>WINDOWS FILE SYSTEM</b>	Searches MS Windows folders for non-searchable content in both Active Monitoring and Backlog modes Searches for image-based PDF, JPG, TIF, PNG, BMP, and emails with attachments Save as Replace Original or New Document, as well as to a New Location	