

# Content Crawler for OpenText

Make every document retrievable and searchable

Improve access to business-critical information

Reduce non-compliance and non-discovery risk

Increase efficiency with automated workflows

Maximize the value of investment in DMS and search technology

Integration with OpenText eDOCS DM and OpenText Content Server

Integration with MS Windows file systems

Access to information in the digital age is crucial. Businesses have invested heavily in Content Repositories such as Document Management Systems as well as in search technology to ensure they have instant access to business-critical documents. Despite this investment, 20% of documents in Content Repositories may be non-searchable and therefore “invisible” to search technology.

## THE RISKS ARE GREAT

Failure to locate a business-critical document can undermine efficiency and productivity as well as put your organization’s reputation and financial well-being at risk when it cannot comply with discovery requests.

## THE SOURCES ARE MANY

Image-based files such as faxes, image PDFs and scanned documents often get profiled in the DMS, or saved to a MS Windows folder through a variety of workflow loopholes; email attachments, legacy documents, mobile technology, documents ingested from acquisitions and imported litigation files. These documents are “invisible” to your search technology.

## THE SOLUTION IS SIMPLE

pdfDocs Content Crawler makes documents retrievable and searchable, reducing the risk to your organization.

*“We use Content Crawler 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**

*“Content Crawler 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**

## SOURCE OF NON-SEARCHABLE CONTENT

- Scanned images saved as TIF or image PDF
- Emails with TIF or image-based PDF attachments
- Electronic faxes saved as TIF or PDF
- Legacy image, PDF or email documents from business acquisitions or litigation file ingestion

## RISKS OF NON-SEARCHABLE CONTENT

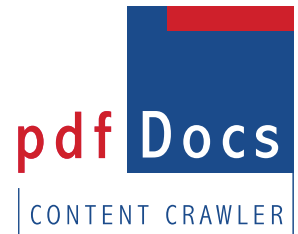
- Non-discovery of critical documents for a case, project, matter
- Failure to comply with Court orders to produce documents
- Productivity loss searching for missing documents
- Investment in DMS and search technologies not being maximized
- User confidence in the DMS system is lost when content is not found

## RETRIEVABLE AND SEARCHABLE

pdfDocs Content Crawler is a framework for searching Content Repositories such as Windows folders, an entire OpenText eDOCS DM or OpenText Content Server database or a subset of documents based on specific queries.

Content Crawler identifies non-searchable content, converts it to a text-searchable PDF using DocsCorp OCR technology and saves it back into the Content Repository.

For OpenText, documents identified as being image documents are saved as either New Versions, Attachments, Related Documents or New Renditions. A text layer is added to the document to facilitate search.



## OPENTEXT

THE CONTENT EXPERTS

OPENTEXT INTEGRATION  
OpenText eDOCS DM  
OpenText Content Server

### SYSTEM REQUIREMENTS OPERATING SYSTEMS

Windows XP Professional (SP3) 32-bit  
Windows 7 32 and 64-bit  
Windows 2008 R2 64-bit  
MS .NET Framework 3.5 and 4.0 Extended



SYDNEY  
MELBOURNE  
LONDON  
NEW YORK  
WASHINGTON DC  
PORTLAND

[info@docscorp.com](mailto:info@docscorp.com)  
[www.docscorp.com](http://www.docscorp.com)

### FLEXIBLE CONFIGURATION

Create any number of definable 'Services' to gain access to a Content Repository  
Assess the text searchability of documents in the Content Repository  
OCR documents that meet the text searchability threshold  
Add a layer of hidden text to a PDF, saving it back into the Content Repository

### DMS SEARCH

Identify image documents, image PDFs and email message files in OpenText by default  
Supports TIF, JPG, PNG and BMP image types  
Caters for multiple databases or libraries

### WINDOWS FILE SYSTEM

Searches MS Windows folders for non-searchable content  
Searches for image-based PDFs, JPG, TIF, PNG, BMP and email messages

### ASSESS TEXT-SEARCHABILITY

Checks all non-searchable content such as image files, image PDFs and emails with attachments  
Identifies PDFs with little or no text. Text-based PDFs will not be processed as they are already text-searchable  
Checks emails stored in a Content Repository to assess text-searchability of the attachments  
Replaces the email attachments with OCR'd PDFs if appropriate

### MAKE SEARCHABLE (OCR)

Documents are processed using OCR technology to generate a new PDF with a hidden text layer  
No requirement for a text file separate to the image or PDF file  
The text layer is searchable  
Use search feature in Adobe Acrobat or Reader to find and review content

### SAVE TO DMS

Integrates with OpenText eDOCS DM 5.1.05 and higher, and OpenText Content Server 9.7.1 and 10  
Uses OpenText API for all connectivity – all security models and privileges honored  
Save the OCR PDF into OpenText eDOCS DM as a New Version, Related Document or as an Attachment  
Save the OCR PDF into OpenText Content Server a New Version or as a New Rendition

### AUDIT AND REPORTING

Rich administrative dashboard to monitor, configure and report on progress  
Maximum control with 'Hold for Review' options prior to OCR and/or Save to Content Repository steps  
Embedded Microsoft SQL database for access to richer reporting if required

### ACTIVE MONITORING

Automate the protection of Content Crawler with the Active Monitoring service  
Assess and OCR newly-profiled or edited document profiles on a regular schedule of your choosing  
Automate workflows to make documents in the Content Repository searchable

Patent pending

pdfDocs Content Crawler and pdfDocs Content Crawler OCR are trademarks of the DocsCorp International Unit. pdfDocs Content Crawler © 2011 DocsCorp International Unit Trust

The application makes use of the following recognition technologies: ABBYY © FineReader © Engine 9.0 © 2008. FINEREADER, ABBYY & ABBYY FineReader are registered trademarks of ABBYY Software Ltd.