IBM

IBM FileNet Whitepaper:
Records Crawler v 3.5
Product Overview

by Karen Pihowich
Records Crawler Offering Manager, Information
Management, IBM Software Group

Introduction

FileNet Records Crawler is a solution for analyzing documents and enforcing document control policies for objects that are not currently under management. It is the first intelligent, rules-based capture and management solution for unmanaged records at the file system level.

Records Crawler seeks out documents that should be better managed, analyzes them using business rules, and either moves them into FileNet for management, or takes some other action on the document, depending on the policies defined by the organization as to what should be done when a relevant document is found. It is primarily intended to do this on documents that are not currently inside a FileNet repository or records management system.

Product Overview

The Business Challenge

Organizations of all sizes require effective records management in order to be compliant with external and internal regulations and achieve business continuity and growth. Hundreds of millions of documents reside uncontrolled on network file share drives, where they are disorganized, unsecured, and often duplicated. With so many users and processes contributing daily to the massive accumulation of business critical documents, enforcing corporate records policies is an increasingly challenging issue.

In today's corporate environment, 80% of documents that should be managed are not. Even after the huge volumes of email generated by companies are brought under control using FileNet Email Manager, hundreds of millions of documents remain uncontrolled on network file share drives. The following are just some of the business drivers behind the need to gain control over unmanaged unstructured data:

- Safety concerns over outdated procedures

- Legal issues arising from poor contract management

- The need to integrate content from multiple systems

- An easier way for users to contribute content

- The desire to recover space on network shares

- The ability to locate and manage records "in place"

The business challenge is how to cost effectively bring documents on the network file share drives under control and ensure they are handled in a manner compliant with corporate and regulatory policies.

Solution Description

FileNet Records Crawler addresses this point of pain for large corporations by enabling a consistent and enforced approach to the management of currently unmanaged records on network file share drives – automatically and cost effectively. Records Crawler is unique in its ability to reduce complexity, risk, time, cost and litigation concerns around data stored on network file share drives. It is the only solution that is:

- **Integrated:** uses rules and profiles pull the File Plan elements from FileNet Records Manager, providing instant records declaration and increased Records Manager automation

- **Enforced:** alleviates end users, records managers and compliance officers of the burden of enforcing compliance and records policies throughout the enterprise

- **Automated:** selectively captures and manages records based on the value of the record, and subsequently automates the entire record lifecycle through FileNet Records Manager for complete records management

- **Simplified:** manages all compliance content rules from a single rules base with simplified compliance profile management

Records Crawler allows organizations to quickly and easily move or manage in place large volumes of files for a multitude of business purposes, including records management and file system management. Using Records Crawler, organizations can easily gain control over the unmanaged content in file systems and network drives, thereby improving adherence to records management and compliance requirements.

Records Crawler provides automated, real-time loading of large volumes of files from monitored locations on network file shares into FileNet, as well as the automated, real-time creation of large volumes of proxy objects in FileNet for files to be managed in-place on the network file share drives. Because Records Crawler can automatically identify documents on the file share that should be declared as records and can automatically declare them as such, Records Crawler provides organizations with a powerful tool for compliance with document retention policies.

Flexible rules-based document loading and indexing permit FileNet properties – including a Records Manager classification and multi-value properties – to be assigned dynamically to each document based upon metadata taken from the files themselves or accompanying metadata files. Post-processing options include the ability to delete, move or flag the source file, or replace it in the network file share with a Windows shortcut to the file imported into FileNet.

Not Another Search Tool
FileNet Records Crawler is different from search and federation services applications in a number of areas.

FileNet Records Crawler	Search Tools
• Automatically monitors file systems based on policy, rules and profiles. • Once a record is identified by rule (i.e. every spreadsheet with "budget" in title) profiles then determines the action (copy, move, stub, extract, delete, declare record). • Direct integration into Records Manager file plan ensures precise classification. • Places content and records under managed control and enforces security, privacy, compliance and audit policies. • Not "auto" classification. • Can trigger business processes or other FileNetP8 events.	• Search is an event. • You typically know what you are looking for (e.g. *where is that budget spreadsheet from last year?*) • Search helps you find or locate based on text or metadata value. • Search does not cause action.

A Complement to Content Federation Services (CFS)

FileNet Records Crawler provides capabilities that are different and complementary to FileNet Content Federation Services.

FileNet Records Crawler	FileNet Content Federated Services
Migrate Enterprise Content into FileNet or Manage Content in Place without Federation • Automatically monitors file system based on policy, rules and profiles. • Integrates NTFS file systems, Image Services, and Content Services into the FileNet P8 platform. • Once a record is identified by rule, profiles then determine the action (copy, move, stub, extract, delete, declare record). • Direct integration into Records Manager ensures content and records under managed control and enforces security, privacy, compliance and audit policies. • Allows content to be used and managed in place, without moving or migrating it from existing repositories • Trigger business processes or other FileNet P8 events without moving the content.	**Manage Enterprise Content in Place through Federation** • Provides the ability to unify multiple content repositories into a single enterprise catalog. • Integrates Image Services, Content Services and 3rd party repositories into the FileNet P8 platform (CS and NTFS coming spring 2006 with more 3rd party repositories to follow). • Enables customers to search for, manage, and actively use content as native FileNet P8 content – regardless of its actual repository. • Enables Records Manager to declare and lock down records in IS, CS and 3rd party repositories without moving content. • Allows content to be used and managed in place, without moving or migrating it from existing repositories. • Trigger business processes or other FileNetP8 events without moving the content.

Capabilities & Benefits

Records Crawler is a server-based application that runs as a "Windows service" and is configured to monitor one or more source locations. In the 3.5 release, the source locations will typically be network drives that Records Crawler has security privileges to access. An administrator configures the Records Crawler to either constantly, or on a scheduled basis, recursively move through all of the folders comparing documents to business rules. These business rules are based on either the document's file system properties (such as filename or creation date) or the properties of the folder for the document that it finds.

As Records Crawler is based on architecture similar to FileNet Email Manager, there are many similarities in the way rules, profiles and the administration functions operate. Once a document matches a business rule, a profile is applied to that document before it is added to FileNet. A profile defines not only indexing information such as target folder location within FileNet as well as other document properties, but it also defines the policy for the organization as to what should be done with that document when it is located. These policies allow Records Crawler to take action on the document such as delete it, replace it with a shortcut, or leave it alone while still recording its existence.

FileNet Records Crawler features:

FileNet ZeroClick™ File Capture

The dynamic nature of FileNet Records Crawler's import features sets it apart from other solutions, and permits sophisticated and targeted file handling. With full support of FileNet ZeroClick™ functionality, FileNet Records Crawler provides a fully automated and centralized document capture process, in which files matching specific business rules are automatically indexed and archived in the FileNet repository. The inclusion of an advanced rules engine greatly enhances the flexibility with which documents can be loaded into FileNet repositories or managed in place in the file share.

On-Demand Processing and Securing of Files

The dynamic nature of Records Crawler's import features sets it apart from other solutions. FileNet Records Crawler can automatically load documents into FileNet as they are placed in a monitored location on the file share, or documents can be loaded following a pre-defined schedule. Records Crawler can replicate the source folder structure in FileNet (in FileNet P8 repositories only), based on the originating folder name, and can even enable folder names to be set dynamically based upon fixed and/or variable metadata pulled from source documents.

Post-processing options include the ability to leave documents in place, move them, flag them, or replace source files with FileNet P8 shortcuts so that users will be able to find documents that had previously been loaded. FileNet Records Crawler also supports the ability to change the file system security to "Read Only" to help maintain the document's authenticity for the period prior to destruction. Business users can access the documents that have been processed by FileNet Records Crawler, but cannot modify or delete them.

FileNet Records Manager Integration

Tight integration with FileNet Records Manager (RM) helps organizations adhere to records management policies by simplifying the declaration of files as corporate records. FileNet Records Crawler fully supports FileNet ZeroClick™ functionality to allow the declaration of documents as records during the capture process – including the dynamic assignment of a FileNet Records Manager classification – eliminating the need for a separate declaration step or event trigger and reducing the opportunity for user error.

For organizations implementing a new FileNet Records Manager system, or migrating records from a legacy file system, the issue of adhering to records management policies is no small task. Since FileNet Records Crawler has been architected to support FileNet Records Manager, administrators and records managers can be assured that policies are adhered to when migrating files into FileNet Records Manager.

Business Process Automation

FileNet Records Crawler allows critical documents residing outside FileNet P8 to initiate and play a significant role in business processes. Organizations thus have a means of connecting content not in the enterprise content management repository to business processes throughout the enterprise, providing value and tangible returns above and beyond the initial investment.

The inclusion of an advanced rules engine greatly enhances the flexibility with which documents can be loaded into FileNet or managed in place in the file share. The rules engine permits sophisticated and targeted file handling, including the ability to populate multi-property fields with multiple values, automatically replicate the source folder structure in FileNet P8, and automatically assign folder or file names based upon document metadata. Records Crawler can automatically load documents into FileNet as they are placed in a location monitored by Records Crawler, or documents can be loaded following a schedule.

Extend Records Management Control

Becoming compliant with external and internal regulations requires many organizations to extend records management control to documents scattered throughout their organizations. Rather than moving huge volumes of files into centralized repositories, FileNet Records Crawler can be used to extend control over organizational file shares and manage documents "in place."

Records Crawler has the ability to change the file system security to Read Only at the time the proxy is created to help maintain the documents' "authenticity" for the period prior to destruction. Business users can access the documents, but cannot delete or modify them.

Bulk Document Import

Organizations often have to import large volumes information into their FileNet repositories, such as statements, invoices, or activity reports, on a frequent and regular basis (e.g. weekly or monthly). FileNet Records Crawler permits automated, scheduled or constant bulk loading of such files into FileNet Content Manager (CM), Image Manager (IM) and Content Services (CS) repositories, so that management processes can be initiated without delay. End users who create large amounts of new content can quickly add multiple files into FileNet with a simple drag & drop, without having to manually fill out properties, saving considerable time and effort. Because the profiles for loading files are pre-defined, FileNet Administrators can be certain users are adhering to established policies.

Metadata File Support

Records Crawler supports both CSV and XML metadata files. Metadata files can be created manually or generated by other applications. Files without associated metadata can still be imported into FileNet using Records Crawler.

High-Volume Throughput

Depending upon network traffic, configuration, file type and size, a single Records Crawler server can import between 25,000 and 40,000 records an hour.

Business Scenarios

Though Records Crawler deployments typically fall into one of the following scenarios, the flexibility of the product has resulted in many variations on the scenarios presented here.

Move Documents off the Network into FileNet
Benefit: The ability to quickly migrate some or all documents not currently under management into FileNet.
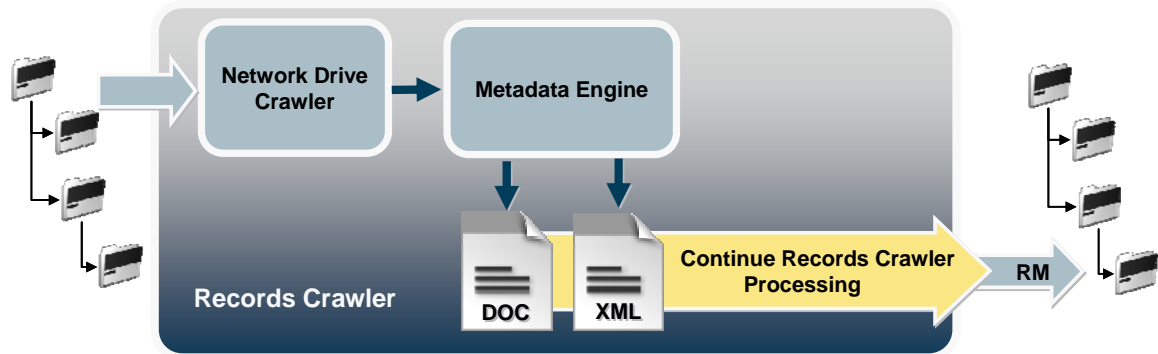
Ideal for: Document migration and compliance.

Figure 1. Moving documents off the network into FileNet using Records Crawler

In this scenario, a customer needs to move large quantities of documents into the FileNet Content Manager and possibly declare them as records en route. Records Crawler provides for a simple configuration that will allow documents to be quickly ingested into FileNet Records Manager (RM), CM, CS, or IS. Records Crawler can replicate the source file system directory structure in the FileNet document repository if desired and even map the source folder hierarchy to the records categories if required (available for FileNet P8 only). The source documents can be deleted, moved or left in place once they have been processed by Records Crawler.

Move Documents and Leave a Link on the Network

Benefit: Provide document management while enabling users to locate documents from network locations.

Ideal For: Extending RM control to documents in file share while leaving them available for users or business processes.
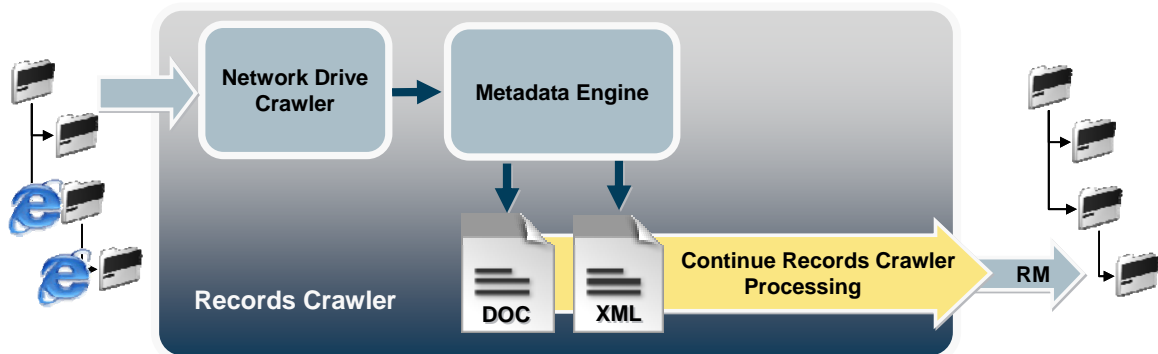
Figure 2. Moving documents into FileNet and leaving a link behind

This scenario is similar to the one above, in which documents on network file share drives in use by users can be moved into FileNet, but FileNet P8 shortcuts can be left behind on the network file share drive so that when a user returns to that network folder to find the document, they will be able to access it from FileNet via the network location in which they expect to find it.

Automation of Business Transactions

Benefit: Intelligent bulk loading of documents and transaction properties into FileNet as a regular business transaction.

Ideal for: Monthly statement imports and automated loading of document batches.

This scenario is common not only for migration efforts, but also for the ongoing movement of large volumes of documents such as statements generated by third-party systems. Using rules and regular expressions, Records Crawler customers import batches of documents into a repository on a daily basis. Rules are combined with "regular expressions" queries to extract account information from the file name and apply it to FileNet properties, ensuring that every statement is linked to its corresponding customer records.
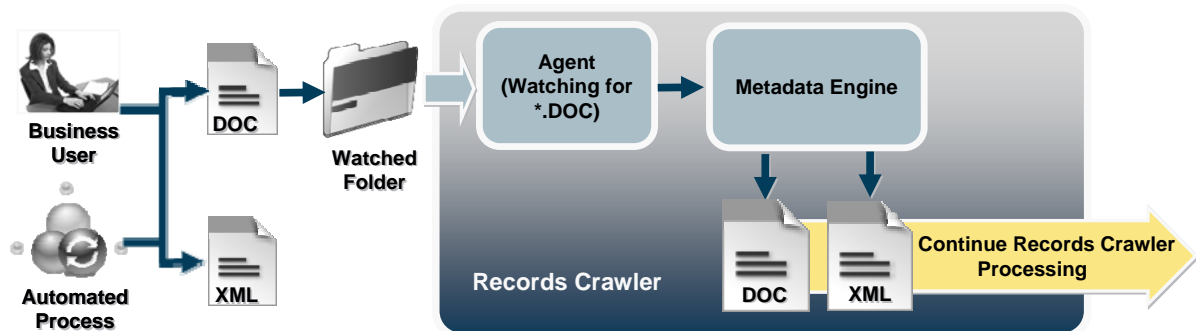
Figure 3. Records Crawler for automation of business transactions

This diagram represents a scenario in which external CSV or XML metadata files are used to define how documents should be added to the FileNet repository. This is of great use for organizations that outsource scanning or document preparation where disks of documents are received and need to be imported. For example, outsourced engineering design work will typically be sent into the company on a CD containing hundreds of CAD files and a CSV "spreadsheet" describing each of the documents.

Simplified User Based Loading

Benefit: Alleviates the need for users to define document properties and manually load documents into FileNet.

Ideal for: Automating the irregular loading of documents into FileNet.

Records Crawler allows end users who create new content to quickly add multiple files into FileNet with a simple drag and drop, without having to manually fill out properties, saving considerable time and effort. Because the profiles for loading files are pre-defined, FileNet Administrators can be certain users are adhering to the file plan.
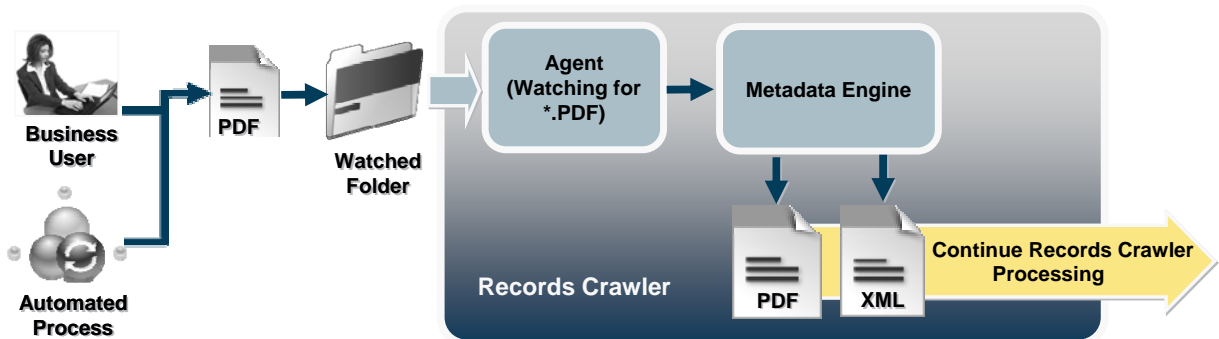
Figure 4. Records Crawler for simplified user based loading of files

In this scenario, a user or an automated process creates only a source file (without external metadata) and places it in a designated folder location on the network file share (e.g. "Contracts"). Records Crawler will retrieve the source file and pass it on to the metadata engine which can produce file system and path information to include in the metadata properties within FileNet (i.e. the name of the source folder can be used a metadata property for each document captured from that location).

Extending RM Control to the Network File Share

Benefit: Enables documents on a network file share to be declared as a record without bringing them into a FileNet repository.

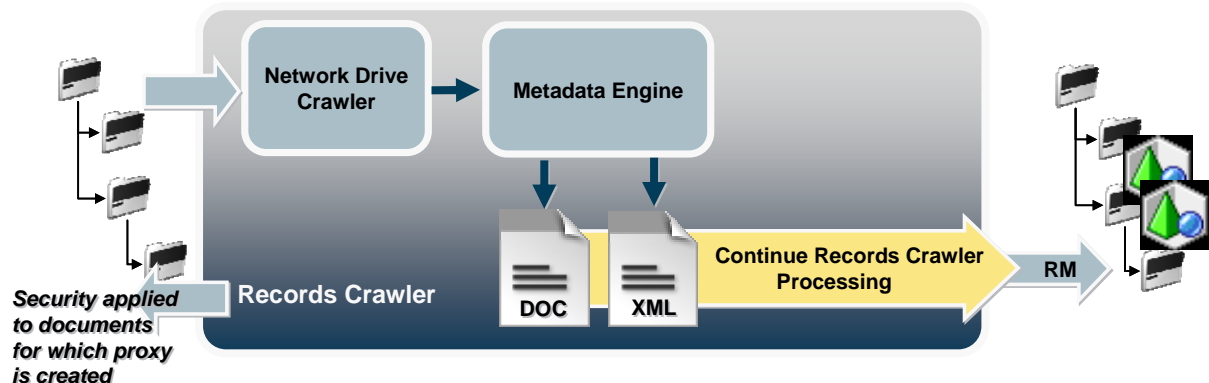**Ideal for:** Reducing the space required to store all file share documents in a repository.

Figure 5. Records Crawler to extend RM control to the network file share

This is a more complex records management scenario whereby an organization wants to manage their documents as records without bringing them into a FileNet repository. The organization wants to leave the documents in the network file system but still manage them as records. In theory, Content Federation Services (CFS) is capable of doing this, but the challenge is in making CFS aware of the documents' existence, identifying which should be declared as records, and performing records declaration intelligently and in bulk.

As Records Crawler moves through the file system and locates documents that match a business rule, the policy is applied that takes note of the document and creates a Record "Proxy Object" in FileNet so that Records Manager is now aware of the existence of that document. In order to maintain the integrity of the document, Records Crawler also changes the security on that document in its source location so that it cannot be altered or destroyed by anyone, except records managers through CFS and RM. These documents are also tagged using a hidden property, so that they're not reprocessed as Records Crawler continues its daily activities.

Business Scenarios: Customer Stories

FileNet customers across a multitude of industries are using Records Crawler to manage their corporate content. Dozens of other prospects are putting Records Crawler through its paces – and are finding time and again that it not only performs as promised (and more), but that it performs better and faster than anything else they've seen. Some highlights include the following:

Major Insurance Company – Support for FileNet Records Manager

After purchasing FileNet Records Manager, this major insurance company needed a way to declare Image Services and Content Services documents as records in FileNet P8, without having to move all the millions of files in those repositories. Records Crawler provided just that ability, as well as the ability to declare documents on the file system as records, while leaving those documents in place. Furthermore, given the success of the Records Manager implementation and Records

Crawler's highly extensible feature set, the company is looking into even more ways of leveraging Records Crawler's capabilities across business functions throughout the organization.

Electric Utility – FileNet P8 migration

A major Canadian electric utility is using Records Crawler to assist in moving drawing records from the existing mainframe computer into FileNet P8. The first phase of the migration involves using Records Crawler to move more than a million drawing records; the second phase of the migration will see Records Crawler transfer approximately 50,000 electronic drawings from the existing GPS system into FileNet P8. This utility was able to use Records Crawler out of the box during the first phase of the migration project, with absolutely no additional configuration required, saving significant time and effort.

Oil & Gas – Bulk loading of scanned and electronic files

An external third party scans oil well files, which are then bulk loaded into the organization's FileNet repository using Records Crawler. A simple spreadsheet updated by the third party with the scanned file names is used as the metadata file. Other files created by end users are added to a network share folder structure monitored by Records Crawler; the folder path is parsed into FileNet properties and the folder structure is replicated in the repository as Records Crawler automatically adds the documents to the FileNet repository.

Financial Services – Bulk capture of client statements and documents

Client statements and documents are generated by an external third party, and require periodic bulk loading in to the central FileNet repository. FileNet document properties – which come from static values, folder names and file attributes – are automatically assigned by Records Crawler as the documents are loaded into the FileNet repository.

Financial Services – Transaction receipts capture into FileNet P8

Instead of using paper transaction slips for deposits, withdrawals and other banking activities, this credit union now captures customer signatures electronically and adds the resulting PDF transaction file into FileNet using Records Crawler. Benefits include improved employee efficiency and reduced costs related to storing, finding and retrieving paper receipts.

Process Manufacturing – Simplified user loading

Employees in a geographically remote office location scan inventory sheets and invoices into watch folders on a network file share, which is monitored on a continual basis by Records Crawler. Records Crawler picks up documents as they appear and adds them automatically to the FileNet repository, saving the administrator and end users time and effort.

Conceptual Overview

FileNet Records Crawler is a server-based application that runs as a "Windows service" and is configured to monitor one or more source locations. The source locations will typically be network drives that Records Crawler has security privileges

to access. An administrator configures Records Crawler to either constantly, or on a scheduled basis, recursively move through all of the folders comparing documents to business rules. These business rules are based on either the document's file system properties (such as filename or creation date) or the properties of the folder in which the document resides.

As Records Crawler is based on architecture similar to FileNet Email Manager, there are many similarities in the way rules, profiles and the administration functions operate. Once a document matches a business rule, a profile is applied to that document before it is added to a FileNet repository. A profile defines not only indexing information such as target folder location within FileNet as well as other document properties, but also defines the policy for the organization as to what should be done with that document when it is located. These policies allow Records Crawler to take action on the document such as delete it, replace it with a shortcut, or leave it alone while still recording its existence.
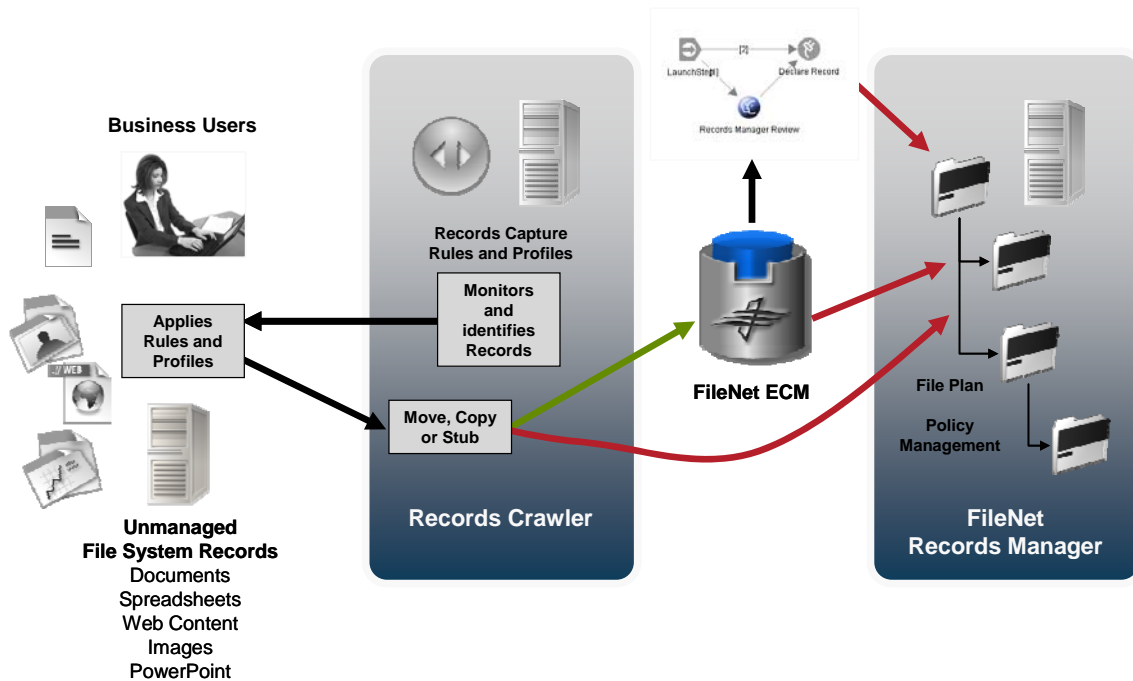
Figure 6. FileNet Records Crawler Workflow

Import Sources

FileNet Records Crawler monitors designated folders – either continuously or on a scheduled basis – for files that are to be imported into FileNet or that are to have a proxy created in FileNet. When files are present, Records Crawler uses metadata files to identify which files are to be uploaded to the FileNet repository and assigns associated properties based

on pre-defined business rules before automatically adding each file to the FileNet system. Records Crawler processes documents from any network drive folder under one of two scenarios:

- **Network Drive Import or File System Management:** Records Crawler is configured to import or create a proxy for the contents of a network drive (including all documents subfolders if desired). Records Crawler assigns associated FileNet properties to each document, based on pre-defined business rules, before automatically adding each file or proxy to the FileNet system. Network drive import is typically to bulk-load documents as part of a system upgrade or bulk import project.

- **Watch Folders:** Records Crawler is configured to monitor one or more folders on a network drive – watch folders – and when documents are placed in these folders, Records Crawler uses document metadata to assign FileNet properties to each document, based upon pre-defined business rules, before automatically adding each document to the FileNet repository. Watch folders are typically employed when Records Crawler is used as part of an ongoing business process.

Metadata Files

Metadata files are simple spreadsheet-like documents (either CSV or XML) created manually by external content contributors or automatically as part of a business process. Once created, the metadata files are simply added to the monitored folders to help Records Crawler determine which files will be added to the FileNet repository. Records Crawler uses metadata files to identify which files in a watch folder or network drive are to be uploaded to the FileNet repository. Records Crawler can also import documents that do not have associated external metadata.

Files with different metadata stored in the same folder location can be handled differently by Records Crawler, depending on specific business requirements. With this configurable flexibility, Records Crawler permits a simple monitored folder structure which requires very little administrative upkeep.

Profiles

Business rules, encapsulated as profiles, are defined by administrators and are used by Records Crawler to populate some or all FileNet properties for all documents moved into FileNet. Profiles determine:

- what data will be inserted in property fields,

- where files will be located in the FileNet system,

- what security settings will be applied to the files, and

- what processing options will be taken on files after they have been imported into FileNet or have a proxy created in FileNet (i.e. delete the file, move to another location, change security setting, etc.).
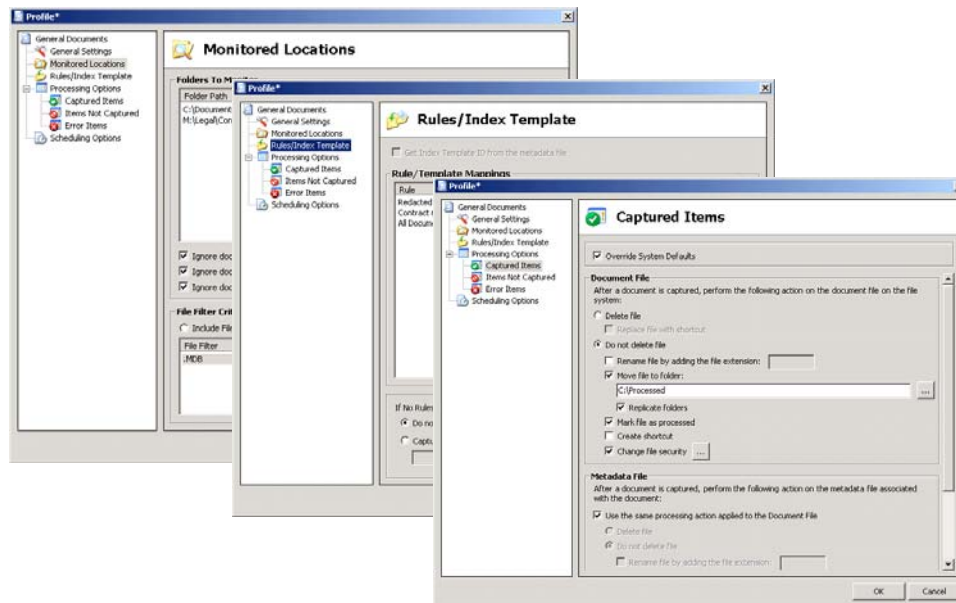
Figure 7. Sample profile configurations

Using profiles, administrators can choose to fill in property fields with static data, information derived from the file system of origin, or a combination of both for documents placed in a system folder.

By creating profiles, Records Crawler administrators can address their organization's unique document import needs. A profile breaks down into the following components:

Configuration Settings
Configuration settings include how the Records Crawler Server is set up, which folders it will monitor (watch folders) or from which network drive and folder(s) it will import documents into a FileNet repository.

Rules
Rules are used to specify what documents should be captured. To enable automatic, rule-based document capture using Records Crawler, the FileNet administrator creates business rules that determine which documents should be automatically captured to a FileNet repository. Rules can be configured based on document metadata, including keywords found in the title or body of the document.
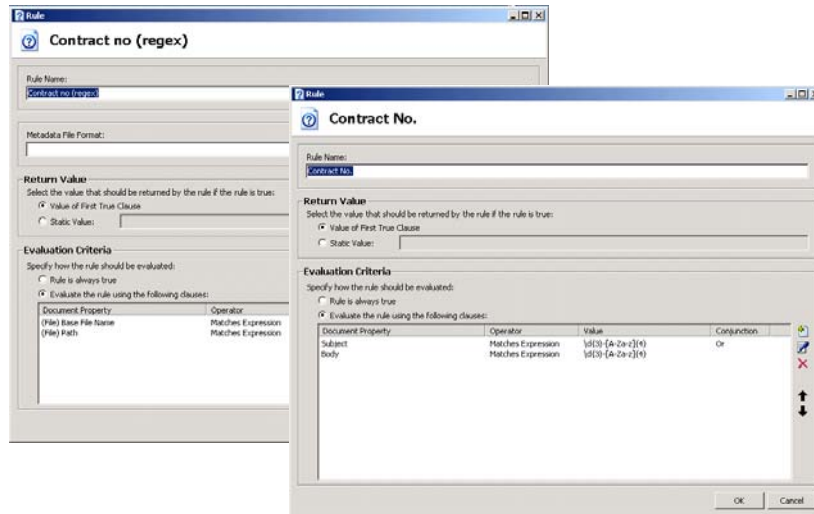
Figure 8. Sample rules screenshots

Documents that do not meet a business rule can be ignored by the Records Crawler Server and not imported into FileNet.

Indexes

Indexes contain the information required to add a document to a FileNet repository, including the name of the repository, how to index the document in the repository (e.g. title) and what security should be applied to the document in the repository. The index also defines whether or not the document is to be declared as a record.
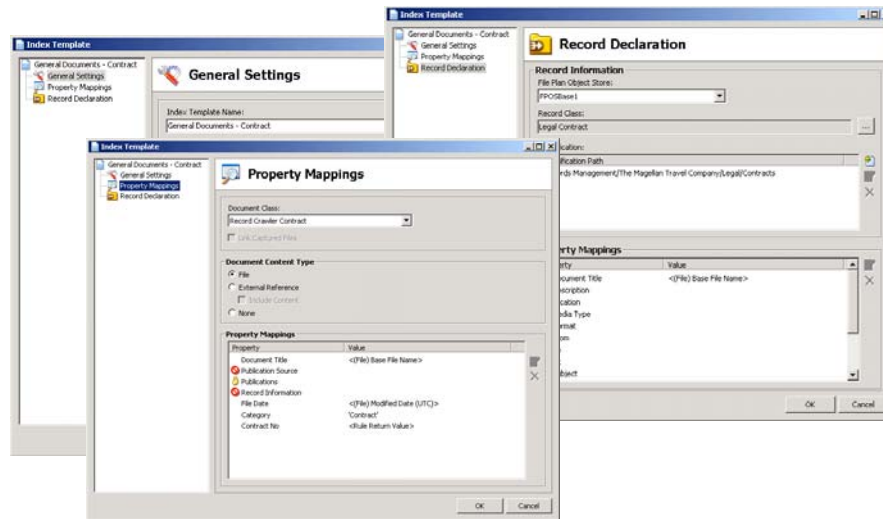
Figure 9. Sample index screenshots

Modes

Records Crawler supports two modes of operation: active mode and scheduled mode. Active mode monitors specified folders – watch folders – and processes new documents against specified rules as soon as they appear in the folder. Scheduled mode processes documents in a specified folder on an administrator defined schedule (e.g. once a day, once a month).

Processing Options

Processing options describe the actions to be taken once a document has been archived in a FileNet repository. Processing options include the option to delete or move a document from the source folder following its capture.

Log Files

Records Crawler also keeps a log of its daily activities by registering dates, times, and outcomes of attempted file imports. If an error occurs during processing, the error code and a detailed description are also noted in the log for fast troubleshooting.

Architecture Overview

Records Crawler is an extensible platform that can analyze documents using business rules and to take action on documents that match those rules in order to enforce corporate policies for document management. It is primarily intended to do this on documents that are not currently inside a FileNet repository or records management system, however, version 3.5 provides optional capabilities for working with select FileNet source repositories.
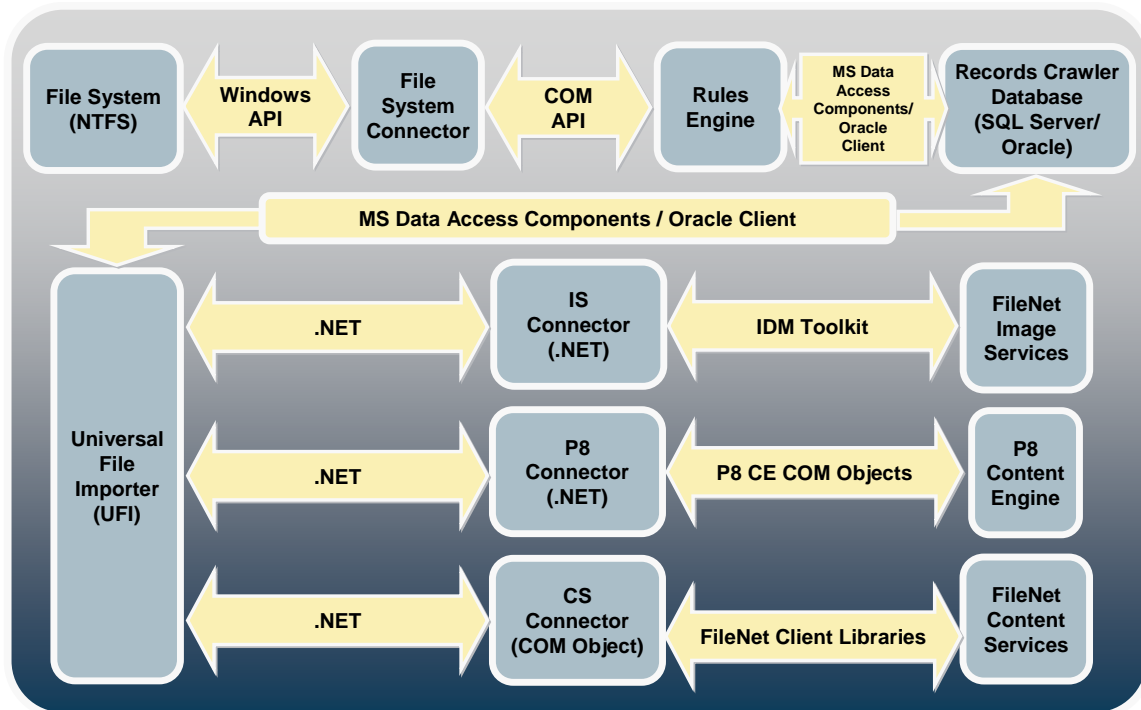
Figure 10. Records Crawler Architecture

As an extensible platform, Records Crawler consists of the following components:

Universal File Importer (UFI)

The Universal File Importer is the transaction engine at the heart of the Records Crawler application. UFI contains all configuration information and processing logic. It is developed using .NET, Java and C-sharp programming languages, as are the other components of the Records Crawler application.

Source System Connectors

These provide the bi-directional "hook" into the system from which documents are being processed. Each source system type requires its own connector. Source connectors are typically written to use any APIs available for the source system, but sometimes direct integration is required to enable specific features. The following source system connectors are available for RC 3.5:

Source System Connector	Availability
• NTFS	• Supported out of the box.
• Content Services	• Available and supported only as part of a custom PS engagement requiring a Statement of Work (SOW). Limited feature set available.
• Image Services	• Available and supported only as part of a custom PS engagement requiring a SOW. Limited feature set available.

Target Systems Connectors

These provide the bi-directional "hook" into the target FileNet document repositories. Target systems connectors use the appropriate FileNet supported APIs (e.g. in Records Crawler 3.5, the COM API is used to put records into Content Manager and the JAVA API is used to query information from Content Manager). Records Crawler provides target system connectors for all FileNet repositories and support for Records Manager:

Target System Connector	Availability
• Content Manager 3.0 and 3.5	• Supported out of the box
• Records Manager 3.0	• Supported out of the box
• Content Services 5.2, 5.3 and 5.4	• Supported out of the box
• Image Services 3.6 and 4.0	• Supported out of the box

Configuration Database

A database required to store configuration settings defined in UFI. This database is not large (under 10MB) and can be co-located with an existing database on a machine other than the Records Crawler server itself (as long as the necessary domain permissions are granted to the Records Crawler server). Records Crawler supports the following databases:

Configuration Database	Availability
• SQL Server 7.0 or 2000	• Supported out of the box
• Oracle 8i and 9i	• Supported out of the box

Server Hardware Requirements

The following are the minimum hardware requirements for Records Crawler:

- Pentium 4, 2.0 GHz

- 512 MB RAM

- 100 MB disk space

Server Software Requirements

The server on which Records Crawler is to be installed requires the following software:

- Windows 2000 Pro, Windows 2000 Server, Windows XP Pro, or Windows 2003 Server

- .Net Framework 1.1

- If Oracle is being used as the database server, one of:

- Oracle 8i Client
- Oracle 9i Client
- Oracle 10g Client

- If FileNet Content Services is being used as the document repository, one of:

- Desktop 3.x and client libraries
- Web 3.x and client libraries

- If FileNet Image Services is being used as the document repository:

- Web 3.x and client libraries

- If FileNet Content Manager is being used as the document repository:

- **Microsoft Advance Directory (AD). Novell eDirectory and SunONE will be supported in Q1, 2006. Check with Product Management for details.**
- FileNet Content Manager COM SDK
- FileNet Client Connectivity

- If FileNet Records Manager is being used:

- **Microsoft Advance Directory (AD). Novell eDirectory and SunONE will be supported in Q1, 2006. Check with Product Management for details.**
- Java Platform Standard Edition 1.4.2

Technical Considerations

The following technical considerations are relevant to the implementation of Records Crawler:

- A single Records Crawler server can connect to a single repository (and can connect to multiple object stores within a single repository).

- If FileNet Content Manager is being used as a document repository, the server upon which Records Crawler is installed must be on the same domain as the Content Manager server.

- As a Windows Service, a Windows domain user account is required if Records Crawler will be accessing network file share drives on other computers. The domain user account (e.g. Records Crawler service) requires the following attributes:

- Must be a member of the Domain Users group
- Must have "Full Control" on any folders on the network file share from which Records Crawler will be capturing document
- Must have "Full Control" on all documents that will be processed by Records Crawler

- As a Windows Service, Records Crawler requires that a single administrator level account be created to enable it to connect to a FileNet document repository. The user created must have the following attributes:

- Content Manager

  - o Connect to store
  - o Create new objects
  - o Modify existing objects
  - o Delete objects

- Image Services

  - o Ensure the user has permissions to add items to the library

- Content Services

  - o Add Item privileges

  - o Author privileges for all folders in the system

  - o Member of Admin group

  - o May Create Document privileges on all document classes required for document capture

Configuration Reference

This section provides an overview of the features and options that can be configured in Records Crawler. More comprehensive information on the available features and options can be found in the **FileNet Records Crawler Configuration Manager Guide**.

All configuration is performed using the Records Crawler Configuration Manager. The main screen of the Configuration Manager consists of the two sections:

- A tree structure that shows the various components that make up Records Crawler, and

A content pane that shows details for the component selected in the tree. Components may be added, removed, edited or copied. For each component, the options available are shown in the content pane, in the right click menu and in the file menu.
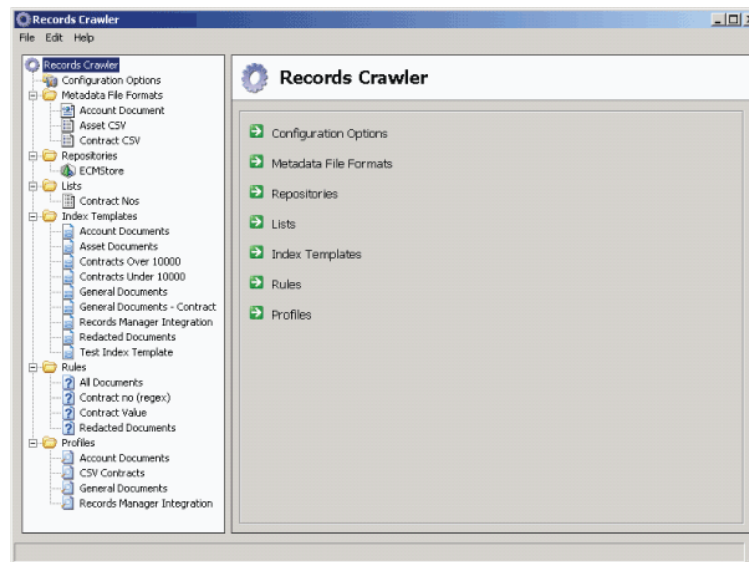
Figure 11. Records Crawler Configuration Manager

Components included in the tree include:

- **Configuration Settings:** contains global system settings for General Settings, Logging and Default Settings.

- **Metadata File Formats**: definitions of metadata files that will be used with Records Crawler.

- **Lists:** List of words that may be used in rule clauses.

- **Repositories:** repositories registered for use with Records Crawler.

- **Index Templates:** templates containing indexing and security information that will be used to add documents to repositories.

- **Rules:** rules to specify what items should be captured.

- **Profiles:** define what locations on a local drive or network share to monitor, what rules to apply to items in the location and what index template to use if a particular rule applies.

The options and settings available for each of these are detailed in the following sections.

Configuration Settings

The Configuration Settings section contains global system settings. The configuration settings consist of:

- General Settings

- Logging

- Default Settings

- Client Configuration

General Settings

The General Settings section contains general settings for the system.

Logging

The Logging section contains settings to enable/disable system logging and to set the type of logging to use.

Field	Description
• Log Type	• Valid values are "Logging Disabled," which turns off logging and "Log To File," which will turn logging on. If "Log To File" is selected, a log file will be created for each system component. The temporary folder location defaults to the "Temp" folder in the application installation folder.
• Log Level	• The type of data that should be written to the log file. Options available include: • **Fatal Errors:** The minimum logging detail. An event is written to the log whenever a severe problem or critical condition occurs. • **All Errors:** An event is written to the log whenever an error condition occurs—such as when a connection attempt to a server fails. • **Warnings:** An event is written to the log whenever a warning condition occurs—such as when the server cannot understand a communication sent to it. • **Information:** An event is written to the log with every significant action that takes place—such as when a document is captured. • **Trace:** The most verbose logging. This option is useful only for debugging purposes. Events are written to the log at individual steps within each process or task, to pinpoint problems. This option should be used with caution as log files may grow to a very large size in a short period of time.
• Log File Location	• The file folder in which log files created by the application are written.

Default Settings

The Default Settings section contains settings that all profiles will use by default. The settings may be overridden in each profile.

Processing Options/Captured Items

The Processing Options for Captured Items section is used to configure settings that apply to a document if the document is selected for capture. The settings only apply to the document on the file system and not the document that is added to the repository. Separate Processing Options for document files and metadata files may be specified. In addition, Processing Options for files that produce errors may be specified.

Document File

The settings specified in this section will be applied to the document being captured.

Field/Section	Description
• Delete file	• If selected, the file will be deleted from the file system. The file will not be copied to the Recycle Bin, it will be permanently deleted.
• Do not delete file	If selected, the file will be left on the file system. Any combination of the following options may be applied to the document that is left on the file system: • Rename file • Move file to another location • Mark file as processed • Change file security When configuring Profiles, care should be taken to ensure a file will not be processed by the system more than once. For example, if a Profile is continually monitoring a specific location, the file should be either moved to another location or marked as processed.
• Rename file by adding the file extension	• If selected, the file will be renamed by adding the file extension entered to the file name. i.e. if the extension "done" is entered, the file "data.txt" will be renamed to "data.txt.done." It is suggested that the file extension entered be added the exclude list in the File Filter Criteria section in the Monitored Location section in a Profile. This should be done to prevent the system from attempting to capture the renamed file. • If a file already exists with the same name that will be used for the renamed file, a number in parentheses will be added to the filename to make the filename unique.
• Move file to folder	• If selected, the file will be moved to the folder specified. To prevent moved files from being captured again by the system, ensure the folder specified is not being monitored in a Profile. • If a file with the same name already exists in the folder, a number in parentheses will be added to the filename to make the filename unique.
• Replicate folders	• If selected, the full path of the location the file originally resided in will be replicated in the "move" folder. The drive letter in the path or the server name in the path will include in the folder hierarchy

Field/Section	Description
	created.
	• For example, if a file originally resided in the location "c:\documents" and the file is moved to the folder "c:\processed," in the folder "c:\processed," the folder "c\documents" will be created: "c:\processed\c\documents."
• Mark file as processed	• If selected, a property is added to the file to indicate that the file has been processed by the system. In addition, a property is added to the file to indicate that the file has been captured.
	• These properties may be used in a Profile to exclude documents that have already been processed/captured.
	• This option is only available for files that reside on an NTFS file system. The "marking" of the file will not alter the file content, but will add a custom property to the file.
• Change file security	• If selected, the security on the file will be changed to the security specified.
	• See the section Edit File Security for details on editing security settings.

Metadata File

The same settings that are available for Document File are available for Metadata File.

Edit File Security

The Edit File Security screen may be used to specify what security should be applied to a file after being processed by the system.

Processing Options/Items Not Captured

The Processing Options for Items Not Captured section is used to configure settings that apply to an item if the item is not selected for capture.

Document File

The same settings that are available for Processing Options/Items Captured/Document File are available for Processing Options/Items Not Captured/Document File.

Metadata File

The same settings that are available for Processing Options/Items Captured/Metadata File are available for Processing Options/Items Not Captured/Metadata File.

Processing Options/Error Items

Processing Options for Error Items is used to configure settings that apply to an item if an error occurs while processing.

Field/Section	Description
• Rename file by adding the file extension	• If selected, the file will be renamed by adding the file extension entered to the file name. i.e. if the extension "done" is entered, the file "data.txt" will be renamed to "data.txt.done." It is suggested that the file extension entered be added the exclude list in the File Filter Criteria section in the Monitored Location section in a Profile. This should be done to prevent the system from attempting to capture the renamed file. • If a file already exists with the same name that will be used for the renamed file, a number in parentheses will be added to the filename to make the filename unique.
• Move file to folder	• If selected, the file will be moved to the folder specified. To prevent moved files from being captured again by the system, ensure the folder specified is not being monitored in a Profile. • If a file with the same name already exists in the folder, a number in parentheses will be added to the filename to make the filename unique.
• Replicate folders	• If selected, the full path of the location the file originally resided in will be replicated in the "move" folder. The drive letter in the path or the server name in the path will include in the folder hierarchy created. • For example, if a file originally resided in the location "c:\documents" and the file is moved to the folder "c:\processed," in the folder "c:\processed," the folder "c\documents" will be created: "c:\processed\c\documents."
• Mark file as processed	• If selected, a property is added to the file to indicate that the file has been processed by the system. • These properties may be used in a Profile to exclude documents that have already been processed/captured. • This option is only available for files that reside on an NTFS file system. The "marking" of the file will not alter the file content, but will add a custom property to the file.
• Change file security	• If selected, the security on the file will be changed to the security specified. • See the section Edit File Security for details on editing security settings.

Scheduling Options

Scheduling options are used to specify when a "scheduled profile" should run. The options only apply to "scheduled profiles" and not other profile types.

Field/Section	Description
• This profile should run	Specifies type of schedule. Based on the type of schedule selected, different options will appear on the screen. Values that may be selected include: • **Always**: the scheduled profile will run continuously. If an item is added to a location being monitored by the profile, a notification will be sent to Records Crawler. Notifications are only enabled for this scheduling option. • **Once**: the scheduled profile will run only once. • **Daily**: the scheduled profile will run on a daily basis at a specific time. • **Interval**: the scheduled profile will run at a specific time interval. • **Weekly**: the scheduled profile will run on a weekly basis at a specific time. • **Monthly**: the scheduled profile will run on a monthly basis at a specific time.

Metadata File Formats

The Metadata File Formats screen is used to define the file format of metadata files that may be used with Records Crawler. Metadata files may be used to include extra data related to a document that is not available in the basic file properties or to specify what documents Records Crawler should load. XML and delimited file formats are supported.

Edit Metadata Definition

The Edit Metadata Definition Screen may be used to define the format of the data in a metadata file as well as the fields in the metadata file.

Field/Section	Description
• File Format Display Name	• The name that should be used to refer to the Metadata File Format. This label will be used when referencing the metadata file format in Index Templates, Rules and Profiles.
• Format Type	• The format of the metadata. Either XML or Delimited may be selected.

Delimited File Format Options

Field/Section	Description
• Delimiter	• The value used as delimiter in the delimited file format. Either a value may be entered or a pre-defined value may be selected.
• Text Qualifier	• If text in the delimited file is enclosed in specific characters, such as quotation marks, enter or select the value used.
• Multi-Value Delimiter	• If a specific column in the file contains multiple values, enter the delimiter used to separate each value in the column. The delimiter entered should not be the same as the column delimiter specified.
• First row contains labels?	• Select this option if the first row of the metadata file contains labels that define what is in each row of the delimited file. Data in first row will be ignored by Records Crawler if this option is selected.
• References multiple files?	• Select this file if the delimited file contains multiple rows, each of which references a different document file. This option should be selected if the metadata file is being used to instruct Records Crawler what files should be captured.
• Select a sample Delimited File	• A sample delimited file may be used to define the format of the metadata. If a valid file is selected, the Sample File section of the screen is populated with a list of rows in the sample file.
• Sample File	• If a sample file is selected, the Sample File section will contain headings for each column in the sample file. If the "First row contains labels" flag is selected, the column headings shown will be named based on the data in the first row in the file. If the "First row contains labels" flag is not selected, the columns will be named "Column" followed by the number of the column. • To add columns to the metadata mappings, drag and drop the required column into the Metadata Mappings section on the screen.
• Metadata Mappings	• A mapping of fields contained in the metadata file format to Records Crawler properties. • If the Edit or Add options are selected, the Edit Delimited File Metadata Mapping screen will be shown.

XML File Format Options

Field/Section	Description
• Select a sample XML File	• A sample XML file may be used to define the format of the metadata. If a valid XML file is selected, the Sample File section of the screen is populated with the XML file structure.
• Sample File	• If a sample XML file is selected, the Sample File section will contain all the elements in the XML file. To add an element to the metadata definition, expand the element completely, until the end of tree is reached and the "function" associated with the tag shows. For most tags the name of the function is "text()." The function, along with the names the elements in the path is called an XPATH. The XPATH is used by Records Crawler to identify the location of an item in an XML file. Select the function and drag and drop the function for the element into the Metadata Mappings section on the screen. The XPATH will be added to the Metadata Mappings.

	• For example, for the XML:
	• <document> <title>the title</title> </document>
	• The XPATH to the element <title> is "/document/title/text()"
• Metadata Mappings	• A mapping of fields contained in the metadata file format to Records Crawler properties. An XPATH is used to define the location of the field in the XML file.

Edit Delimited File Metadata Mapping

The Edit Metadata Mapping screen may be used to define or update specific details related to a metadata mapping.

Field/Section	Description
• Column Number	• The location of the column in the metadata file. The location is based on the number of the column.
• Display Name	• The display name to use to refer to the metadata property. This is the name that will be used to refer the property throughout the application.
• Data Type	• The data type of the data in the metadata property.
• Date Format	• If the metadata property is a date property, format of the date in the metadata file. To specify the date format, used the following tokens: • M = Months as 1 – 12 • MM = Months as 01 – 12 • MMM = Months as Jan – Dec • MMMM = Months as January – December • d = Days as 1 – 31 • dd = Days as 01 – 31 • ddd = Days as Mon – Sun • dddd = Days as Monday – Sunday • y = Years as 1, 2, … , 99 • yy = Years as 00 – 99 • yyyy = Years as 1900 – 9999 • h = Hours as 0 – 12 • hh = Hours as 00 – 12 • H = Hours as 0 – 23

Field/Section	Description
	• HH = Hours as 00 – 23 • m = Minutes as 0 – 59 • mm = Minutes as 00 – 59 • s = Seconds as 0 – 59 • ss = Seconds as 00 – 59 • tt = AM/PM • t = A/P
• Date is in UTC format	• If the date is a UTC date, select this flag. • This option only applies to FileNet P8. Because all dates in P8 must be stored in UTC format, if the date specified is not in UTC format, it will be converted to UTC format before adding the date to P8.
• For use in rules	• If selected, the metadata will be available for use in rules.
• For use in index templates	• If selected, the metadata will be available for use in index templates.
• Custom system designator	• If the metadata contains system specific data such as a file name or index template reference, select a system designator. • If the metadata property is marked as being a "filename" property and Records Crawler is configured to determine the name of the document file to capture based on data in the metadata file, the marked metadata property will be used. • If the metadata property is marked as being an "index template" property, Records Crawler may be configured to base the Index Template on the value in the metadata property. The value must be an index template ID (which can be obtained from the content pane by selecting an Index Template in the Index Template tree).

Edit XML File Metadata Mapping

The Edit Metadata Mapping screen may be used to define or update specific details related to a metadata mapping. The options are the same as for Delimited File Metadata Mapping.

Lists

Lists may be used to define lists of data, such as synonyms for words, for use in Rules. Rules can be created that will look for a word or phrase that appears in a list in a document property.

Edit List

The Edit List screen may be used to define a list of values.

Field/Section	Description
• List Name	• The name to use to refer to the List.
• Description	• A description of the list content (optional).
• List Values	• Values contained in the list. To add a value, click the Add button. The order that items appear in the list is the order in which the items are evaluated when the list is used in a Rule.

Edit List Value

Field/Section	Description
• Value	• The value of the list item. This is the value that will be used when evaluating a rule clause.
• Return Value	• The value returned to the rule if a value matches criteria defined in a rule. The return value may be used with Index Templates to dynamically assign values in Property Mappings. • The Return Value will default the same text entered in the Value field.
• Description	• An optional description of the value.

Repositories

The Repositories section contains document repositories that have been registered for use with Records Crawler.

Repository Connection

The fields shown on the Repository Connection screen will vary based on the repository type being used.

Content Manager Repository Connection

For a Content Manager Object Store, the following fields are displayed:

Field/Section	Description
• Object Store	• The object store that should be registered for use with Records Crawler. A list of all object stores available on the domain will be shown.
• Repository Name	• The name that should be used to refer to the repository in Records Crawler. The name defaults to the display name of the selected Object Store, but may be changed to any name.
• User	• The domain user that Records Crawler should use when connecting to the object store to add documents. To select a domain user click the Find button.
Password / Confirm Password	The password for the User.

Content Services Repository Connection

For a Content Services Library, the following fields are displayed:

Field/Section	Description
• Library	• The library that should be registered for use with Records Crawler.
• Repository Name	• The name that should be used to refer to the repository in Records Crawler. The name defaults to the display name of the selected library, but may be changes to any name.
• User	• The CS library user that Records Crawler should use when connecting to the library to add documents.
• Password / Confirm Password	• The password for the User.
• Group	• The CS group that the user should utilize when logging in to the library.

Image Services Repository Connection

For an IS Library, the following fields are displayed:

Field/Section	Description
• Library	• The library that should be registered for use with Records Crawler. A list of all libraries available on the domain will be shown.
• Repository Name	• The name that should be used to refer to the repository in Records Crawler. The name defaults to the display name of the selected library, but may be changes to any name.
• User	• The IS library user that Records Crawler should use when connecting to the library to add documents.
• Password / Confirm Password	• The password for the User.

Index Templates

The Index Templates section contains index templates that have been created for use in Profiles. The index template specifies what document repository an item should be stored in as well as what properties and security settings should be assigned to the item.

Index Template

The Index Template consists of: General Settings, Property Mappings, Record Declaration (Optional), Permission Mappings, and Client Access Security (if the Index Template is for client use).

General Settings

Field/Section	Description
• Index Template Name	• The name of the Index Template. The name must be unique.
• Metadata File Format	• The metadata file format that will be used in the index template. Metadata properties associated with the Metadata File Format will be available for use in the Property Mappings section of the Index Template. • Selection of Metadata File Format is optional.
• Repository	• The repository that documents should be added to. Only repositories that have been registered for use with Records Crawler are available for use.
• Repository Folder to Add Item	• The folder in the selected Repository that documents should be filed in. Selecting a folder is optional. • The folder may be a static value (a pre-existing folder) or the folder name may be dynamically assigned.
• Declare as Record	• If selected, the section Record Declaration will be enabled to allow record declaration options to be configured. This option is only available for CM 3.0 repositories that have Records Manager installed.

Select Folder

The Select Folder screen may be used to either select an existing folder to file documents into or, if CM is being used, to specify dynamically folder assignment.

Field/Section	Description
• Select Existing Folder	• If selected, the screen will show in Select Existing Folder mode, which will allow an existing folder in the document repository to be selected.
• Advanced Folder Definition	• If selected, the screen will show in Advanced Folder Selection mode, which will allow the folder name to be based the value of a document property, static value, calculated value or regular expression applied to a document property. • This option is only available if CM is being used.

Property Mappings

The following property mapping settings are available:

Field/Section	Description
• Document Class	• The document class to use when capturing documents. All document classes for the repository selected in the General Settings are listed.
• Property Mappings	• A list of Properties (including custom properties) configured for the Document Class selected. The Property Mapping Value screen will show to allow a value to be specified.

Property Mapping Value

The Property Mapping Value screen may be used to assign a value to a repository property. A document property may be selected, a literal value may be specified, a calculated value may be specified or a regular expression may be specified.

Field/Section	Description
• Document Property Value	• Select a document property that should be mapped to the repository property. Only document property that are compatible with the data type of the repository property are shown. • If a Metadata File Format has been selected in the Index Template, metadata associated with the Metadata File Format will be shown in the list of document properties. Standard file properties will always show in the document properties list regardless of the Metadata File Format selected.
• Literal Value	• Enter or select a static value. Only values that match the data type of the repository property may be entered.
• Calculated Value	• A concatenation of document properties and static values
• Regular Expressions	• A value may be extracted from a document property by applying a regular expression to the document property.

Edit Calculated Value

The Edit Calculated Value screen is used to edit calculated values. Select the values and operation to be performed between each value.

Field/Section	Description
• Value	The type of value entered is dependent on the data type of the repository property the calculated value is being assigned to. For strings, a value may be one of: • string value • document property For numbers, a value may be a number value. • For dates, a value may be: • number value • date document property For dates, one document property must be selected and one number value must be entered.
• Operator	The operator displayed is dependent on the data type of the repository property the calculated value is being assigned to. • Strings:

Field/Section	Description
	• + : will concatenate values together
	Numbers:
	• + : will add values together
	• : will subtract values
	• : will multiply values
	• / : will divide values
	Dates:
	• + Day: will add the number of days specified to the date property
	• + Month: will add the number of months specified to the date property
	• + Year: will add the number of years specified to the date property
	• Day: will add the number of days specified to the date property
	• Month: will add the number of months specified to the date property
	• - Year: will add the number of years specified to the date property

Edit Regular Expressions

The Edit Regular Expression screen allows a regular expression to be created.

Field/Section	Description
• Document Property	• The document property the regular expression should be applied to.
• Regular Expression	• The regular expression to apply to the document property selected. Standard regular expression syntax may be used.
• Replacement String	• The match or sub-match to extract from the regular expression. A match is defined in the regular expressing by enclosing the relevant part of the expression in parentheses.
• Default Value	• In the case the regular expression entered does not match any text in the document property selected; the value entered in Default Value will be used.
• Validate	• Click the Validate button to check whether the value entered in the Regular Expression field is valid.
• Test Value	• Enter a test value to which the Regular Expression entered will be applied to when the Test button is clicked.
• Result	• The result of applying the Regular Expression and Replacement String entered to the Test Value entered.
• Test	• Click the Test button to apply the Regular Expression and Replacement String to the value entered in Test Value. The resulting match will appear in the Test Result field. If there is no match, the text "No Match" will appear.

Record Declaration

The Record Declaration section allows properties related to declaring a record in Records Manager to be configured.

Records Manager must be installed and configured to declare a record.

Field/Section	Description
• File Plan Object Store	• The file plan Object Store that the record should be declared in.
• Record Class	• The record class to use when declaring the record.
• Classification	• The classification(s) that should be used when declaring a record • The classification may be a static value or the may be dynamically assigned (i.e. based on the name of the folder an item is located in).
• Property Mappings	• A list of Properties (including custom properties) configured for the record class selected. A value may be assigned to the property by double clicking next to property or click in the "add" icon. The Property Mapping Value screen will show to allow a value to be specified.

Select Classification

The Select Classification screen may be used to either select an existing classification or to dynamically specify classification assignment.

Field/Section	Description
• Select Existing Classification	• If selected, the screen will show in Select Existing Classification mode, which will allow an existing classification in RM to be selected.
• Advanced Classification Definition	• If selected, the screen will show in Advanced Classification Selection mode, which will allow the classification name to be based the value of a document property, static value, calculated value or regular expression applied to a document property. • The name that is dynamically determined must be the full classification path (i.e. \Records Management\Data\Inventory) and must exist as a classification in RM. The system will not create items in the file plan. If a classification does not exist, an error will be logged by the system.

Select Existing Classification Mode

Field/Section	Description
• Select Existing Classification	• If selected, the screen will show in Select Existing Classification mode, which will allow an existing classification in the file plan to be selected.

Advanced Folder Definition Mode

Field/Section	Description
• Document Property Value	• Select a document property that should be mapped to the classification name.
• Literal Value	• Enter or select a static value to assign to the classification name
• Calculated Value	• A concatenation of document properties and static values that will be used to determine the classification name.
• Regular Expressions	• To determine the classification name, a value may be extracted from a document property by applying a regular expression to a document property.

Permission Mappings

The Permission Mappings section of the Index Template is used to specify the security that should be applied to a document when it is added to a repository. Depending on the type of repository being used, the Permission Mappings section will vary in appearance.

Security privileges may be created based on document repository Users and document repository Groups.

Content Manager Permission Mappings

Field/Section	Description
• Security Grid	• The Document Security grid shows all the access privileges that have been created. The grid shows the Type of the access privilege (User or Group), the Name (either the user name or group name) and the access level assigned.
• Users / Groups Radio Buttons	• When adding Security privileges, select the item that the access privilege will be based on. • Selecting Users will show search fields that will allow a domain user to be searched for. • Selecting Groups will show search fields that will allow a domain group to be searched for.
• Access Levels	• The access level to assign to the item selected (either Group, or User). • For Content Manager, valid access levels are: Owner Control, Promote Version, Modify Content, Modify Properties, View Content, View Properties, and Publish. A combination of these access levels may be selected for each user/group.
• Add	• Click the Add button to add the selected item with the access level specified to the Security Grid.

Content Services Permission Mappings

Field/Section	Description
• Document Security Grid	• The Document Security grid shows all the access privileges that have been created. The grid shows the Type of the access privilege (either, User or Group), the Name (either user name or group name) and the access level assigned (either None, Viewer, Author, Owner or Admin). • To edit an existing security setting, select the item and click the Edit button. To remove an existing security setting, select the item and click the Delete button.
• Users / Groups Radio Buttons	• When adding Security privileges, select the item that the access privilege will be based on. Selecting Users will populate the Name list box with a list of document repository Users. Selecting Groups will populate the Name list box with a list of document repository Groups.
• Access Level	• The access level to assign to the item selected in the Name combo box. • For FileNet Content Services, valid access levels are: None, Viewer, Author, Owner or Admin. Only one of these access levels may be selected for each user/group.
• Add	• Click the Add button to add the selected item with the access level specified to the Security Grid.

Image Services Permission Mappings

Field/Section	Description
• Read	• The user or group "**Read**" permissions should be assigned to.
• Write	• The user or group "**Write**" permissions should be assigned to
• Execute	• The user or group "**Execute**" permissions should be assigned to.

Rules

Rules are used to specify what documents should be captured. In a Profile, each rule is assigned to a particular location monitored by Records Crawler, causing the rule to be applied against any document that is placed in the monitored location. If a document meets a particular rule, the document will be captured according to the Index Template assigned to the rule in the Profile.

Rule

The Rule screen is used to define capture rules. The rules are used to determine what items should be added to the repository.

Field/Section	Description
• Rule Name	• **The name of the rule.** The Rule Name will default to "New Rule." Change the name to something more descriptive that indicates what the rule is for. • The Rule Name is a mandatory field.
• Metadata File Format	• The metadata **file format that will be used in the rule.** Metadata properties associated with the Metadata File Format will be available for use in the rule clause • Selection of Metadata File Format is optional.
• Return Value	• **The value that should be returned by the rule (for use by the Index Template) if the rule is true.** The value may be a static value (static text up to 255 characters in length) or the value may be determined by the rule clause. If the value is determined by the rule clause, the value is based on the value in the first clause that is true. The clause value is dependent on the type of the clause value. If the value is a text value, the clause value is the text value. If the value is an expression, the clause value is the result of applying the expression to the document property used in the clause. If the value is a list, the clause is the return value of the first item in the list that matches the document property value.
• Rules Always True	• Selecting this option will cause the rule to always be true. If this option is selected, clauses cannot be entered.

Clauses

Each rule is made up of one or more clauses. The clauses are used to define the logic that determines whether a document should be captured. Each clause contains a Document Property, Operator, Value and Conjunction.

Edit Clause

When a Document Property is selected, the Operator changes according to the type of Document Property selected.

The Document Property may be any standard file properties or a property contained in the metadata file definition.

The Operator may be one of:

• **Is equal:** only shows if a Document Property of type string is selected.

• **Not equal:** only shows if a Document Property of type string is selected.

• **Like:** only shows if a Document Property of type string is selected.

• **Not like:** only shows if a Document Property of type string is selected.

• **Is after:** only shows if a Document Property of type date is selected.

- Is greater than or equal: only shows if a Document Property of type date is selected.

- Is before: only shows if a Document Property of type date is selected.

- Is less than or equal: only shows if a Document Property of type date is selected.

- **Matches Expression:** if selected, the Value may only be a regular expression. Selecting this operator will apply the regular expression to the Document Property selected to determine if it matches.

The Value may be one of:

- **Value:** shows for all Operators. If the document property value matches the value entered (based on the Operator), the clause is true.

- **List:** shows for all Operators. A list is a list of text values that will be compared to the Document Property value when the rule is evaluated.

- **Regular Expression**: only shows if the operator Matches Expression is selected. The regular expression will be applied to the Document Property. If the regular expression results in a match in the Document Property, the clause is true. To enter a regular expression, click the "…" button. A new screen will open to allow for entry of the regular expression.

The Conjunction may be one of:

- **And:** forms a logical "and" relationship between clauses.

- **Or:** forms a logical "or" relationship between clauses.

At least one clause must be entered. The order in which the clauses appear on the screen is the order in which clauses are evaluated. To change the order of the clauses, use the "up" and "down" buttons on the right side of the screen.

For document to meet the criteria specified in the rule, the rule's clauses must evaluate to "true."

Profiles

Profiles define what locations on the file system Records Crawler should monitor, what rules to apply to items in the location and what index template to use if a particular rule applies. In addition, processing and scheduling options may be configured. Prior to creating a Profile, all Index Templates and Rules required should be pre-created.

Profile

The Profile screen may be used to create or edit a Profile. A Profile includes the sections General Settings, Monitored Locations, Rules/Index Templates, Processing Options and Scheduling Options.

General Settings

Field/Section	Description
• Profile Name	• The name of the Profile. Give the profile a descriptive name that indicates what the Profile is for. • The Profile Name is a mandatory field.
• Active	• If selected, the profile will be enabled. If not selected, the profile will be disabled.
• This profile will monitor for	• What the profile should monitor for. Options include Documents or Metadata Files. If Documents is selected, fields to specify how to determine the location of the associated metadata file will be available.

Metadata File Options

Field/Section	Description
• There is a metadata file associated with each document that will be processed	• Specifies whether the document file is accompanied by a metadata file. If selected, the options to specify a metadata file format and fields to specify how to determine the metadata file name will be enabled.
• Metadata File Format	• The metadata file format that will be used in the profile. Only rules and index templates that match the selected metadata date file format will be available for use in the profile.
• Base the name of the Metadata File on: Document Name + a file extension	• If selected, the metadata file name will be determined by replacing the file extension of the document with the file extension entered. i.e. if the document name is "document.pdf" and the file extension "xml' is entered, the metadata file name should be "document.xml."
• Base the name of the Metadata File on: Regular expression applied to document file name	• Apply a regular expression to the document file name to determine the metadata file name.

Field/Section	Description
• Maximum time to wait for file	• The number of seconds the system should wait for a metadata file associated with a document.

Document File Options

Field/Section	Description
• Metadata File Format	• The metadata file format that will be used in the profile. Only rules and index templates that match the selected metadata date file format will be available for use in the profile.
• Base the name of the document to load on: Data in the metadata	• The metadata file contains the name of the document file. The field in the metadata designated as "File Name" will be used by the system to determine the document name.
• Base the name of the document to load on: Metadata file name + a file extension	• The document file name will be determined by replacing the file extension of the metadata file with the file extension entered. i.e. if the metadata file name is "document.xml" and the file extension "pdf' is entered, the metadata file name should be "document.pdf."
• Base the name of the document to load on: Regular expression applied to Metadata file name	• Apply a regular expression to the metadata file name to determine the document file name.
• Maximum time to wait for file	• The number of seconds the system should wait for a document file associated with a metadata file.

Monitored Locations

The Monitored Locations section may be used to specify what locations on the file system will be monitored for documents.

Field/Section	Description
• Monitored Locations	• A list of locations on the file system that Records Crawler should monitor.
• Ignore documents already processed	• If selected, any documents that have already been processed by a profile where the Processing Option "mark as processed" was selected will be ignored.
• Ignore documents already captured	• If selected, any documents that were captured by a profile where the Processing Option "mark as processed" was selected will be ignored.
• Ignore documents where access is denied	• If selected, if the system does not have access to a document, the document will be ignored.

File Filter Criteria

Field/Section	Description
• Include Files	• If selected, only files with the file extensions entered will be processed.
• Exclude Files	• If selected, files with the file extensions entered will not be processed.
• File Filter	• A list of file extensions to include/exclude.

Edit Monitored Location

The Edit Monitored Location screen may be used to add or edit a monitored location.

Field/Section	Description
• Location	• The location to monitor. The location may be a UNC path or path on the local drive.
• Monitor Sub Locations	• If selected, sub locations of the Location specified will also be monitored. Otherwise, only the "root" of the location will be monitored.
• Folder Depth	• If sub locations are being monitored, the folder depth specifies how many levels of sub locations Records Crawler should monitor.

Rules/Index Templates

The Rules/Index Templates screens may be used to specify what rules should be applied to documents in the Monitored Locations and, if a rules applies, what Index Template should be used to add the document to the document repository.

Field/Section	Description
• Rule / Index Template Mappings	• This section contains a list of rules that should be applied to the Monitored Locations and what Index Template should be used to capture a document if a specific rule applies. • Only Rules/Index templates that match the Metadata File Format of the Profile or do not have a Metadata File Format selected will be available.
• Do not capture item	• If selected, if none of the rules apply, the item will not be captured.
• Capture item using this template	• If selected, if none of the rules apply, the item will be captured using the index template selected.

Processing Options

Processing Options available are identical to the processing options available in the Configuration Options. Each profile, by default, will use the Default Settings for processing options defined in the Configuration Options. The processing options may be overridden by selecting the "Override System Defaults" option.

Scheduling Options

By default, the Scheduling Options defined in the Default Settings in the Configuration Options will be used. The Scheduling Options may be overridden by selecting the "Override System Defaults" option.

Note that the body copy style—which is shown here—will automatically be applied once you hit "Enter" (or, create a hard return) after applying the Heading One style. This applies to all heading and subheading styles in this template. [1]

Additional Resources

The following resources are available to provide a greater understanding of Records Crawler.

FileNet Records Crawler Set Up Guide

The Set Up Guide provides detailed information on the requirements and procedures for installing Record Crawler as well as implementation worksheets.

FileNet Records Crawler Configuration Manager Reference Guide

The Configuration Manager Reference Guide provides detailed information on each of the configuration settings in Records Crawler as well as samples of commonly used regular expressions.

IBM

About IBM ECM

As the clear market leader in Enterprise Content Management (ECM), IBM's ECM solutions help organizations make better decisions, faster by managing content, optimizing business processes and enabling compliance through an integrated information infrastructure. IBM's ECM portfolio delivers a broad set of capabilities and solutions that integrate with existing information systems to help organizations drive greater value from their content to solve today's top business challenges. The world's leading organizations rely on IBM enterprise content management to manage their mission-critical business content and processes.