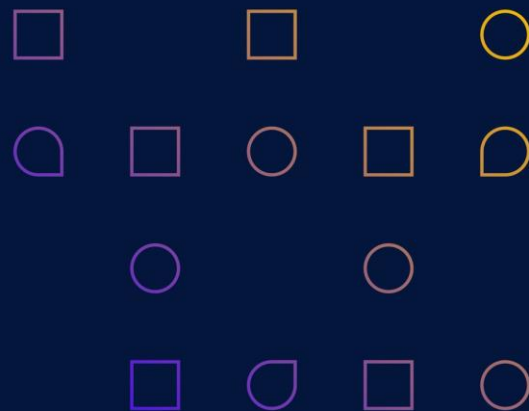


# Discovery Center User Guide

4.17.0 - MIP Stabilization

August 2022



# Contents

- INTRODUCTION..... 11**
- PRODUCT OVERVIEW ..... 12**
  - MODULES AND COMPONENTS..... 12
  - TECHNOLOGIES..... 13
  - SYSTEM PERFORMANCE ..... 13
- CONCEPTS AND PRINCIPLES ..... 14**
  - SECURITY..... 14
    - Roles and Responsibilities*..... 14
    - Access to Information* ..... 14
    - Privileged User*..... 14
  - NETWORK MAP, INDEXES AND AREAS OF INTEREST ..... 15
  - SKIMS, ANALYSIS AND CLASSIFICATION ..... 15
  - CALCULATED FIELDS AND EXTRACTION RULES ..... 16
  - MICROSOFT INFORMATION PROTECTION SENSITIVITY LABELS ..... 17
  - REVIEWING..... 17
    - Using Work Packages* ..... 17
    - Sharing Saved Views*..... 18
- SYSTEM OVERVIEW ..... 19**
  - DISCOVERY CENTER ..... 20
- HOME ..... 21**
  - DISCOVERY CENTER HOME - SYSTEM ADMINISTRATION..... 21
  - DISCOVERY CENTER HOME - AN ADMINISTRATION ..... 21
  - DISCOVERY CENTER HOME - INFORMATION MANAGEMENT ..... 22
  - DISCOVERY CENTER HOME - REVIEWING ..... 22
  - DASHBOARD CHARTS..... 23
  - USER PROFILE ..... 24
    - Notifications* ..... 24
  - APPLICATION LOGS..... 26
- NETWORK MAP..... 28**
  - NETWORK MAP..... 28
    - Populating the Network Map* ..... 30
    - Removing Locations*..... 30
    - Preventing Locations from being Indexed* ..... 31
    - Network Map Management*..... 32
    - Considerations for Credential Configuration* ..... 32
    - Restoring Locations to the Network Map* ..... 33
    - Understanding when Locations are Hidden*..... 33
  - STORAGE TIERS..... 34
- SYSTEM SETTINGS ..... 36**
  - LICENSING ..... 37
    - License Summary* ..... 37
    - Apply License File*..... 38



- System Usage..... 42
- PERMISSIONS AND ACCESS..... 44
  - File system vs Share permissions ..... 45
  - SharePoint Content Selection ..... 45
- USERS AND ROLES..... 45
  - Role Mapping..... 46
  - User Management..... 47
- CREDENTIAL MANAGEMENT..... 49
- DISCOVERY CENTER ..... 51
  - Schedule Constraints..... 52
  - Global Settings..... 53
  - MIP Settings..... 54
- EMAIL CONFIGURATION ..... 56
- METADATA ..... 57**
  - CALCULATED FIELDS ..... 58
    - Fixed Fields..... 58
    - Markup Calculated Fields ..... 59
    - Calculated Field Details..... 59
    - Adding a New Field ..... 59
    - Editing a Calculated Field ..... 62
    - Understanding Matching Strategies and Calculated Field Types ..... 62
  - EXTRACTION RULES ..... 64
    - Adding a New Rule..... 65
    - Editing a Rule..... 67
    - Keyword Match..... 67
    - File Path/Content Pattern Match ..... 70
    - Repository Property ..... 73
    - Embedded File Property..... 73
  - SHAREPOINT PROPERTIES ..... 74
  - OPENTEXT CONTENT SERVER PROPERTIES ..... 75
  - CLASSIFICATIONS..... 75
  - METADATA VALUES IMPORT ..... 76
    - Formatting Multiple Values ..... 76
    - Use of Quotes and Commas..... 76
    - XLSX Date and Time Formats..... 76
- INDEXES ..... 78**
  - ABOUT INDEXES..... 78
    - Overlapping indexes ..... 78
  - INDEXING OVERVIEW..... 78
    - The Indexes list ..... 80
    - Filtering..... 82
    - Sorting ..... 82
    - Index Page Sizes..... 82
    - Volume Under Management ..... 82
    - Index Status ..... 82
  - INDEX CONFIGURATION..... 86
    - Filtering..... 86
    - Index Configuration Page Sizes..... 87



ADDING AN INDEX CONFIGURATION .....	87
<i>Calculated Fields</i> .....	88
<i>Analysis Filters</i> .....	89
<i>Discovery Options</i> .....	91
<i>Analysis Options</i> .....	93
ADDING AN INDEX .....	94
ADDING MULTIPLE INDEXES .....	96
OPTIMIZING AN INDEX .....	97
<i>Optimizing Analysis Performance</i> .....	98
<i>Scheduling an Index</i> .....	98
<b>ACTIVITY .....</b>	<b>100</b>
CURRENT ACTIVITY .....	100
<i>Running Task</i> .....	100
<i>Queued Tasks</i> .....	101
<i>Filtering</i> .....	102
<i>Sorting</i> .....	102
<i>Page Sizes</i> .....	102
ACTIVITY HISTORY .....	102
<i>Filtering</i> .....	103
<i>Sorting</i> .....	103
<i>Page Sizes</i> .....	104
TASK STATUS REPORTS .....	105
<i>'Processing Index' Task Status reports</i> .....	105
<i>'Actions' Task Status Report and Audit Trail</i> .....	106
<i>File Labeling Status Report and Audit Trail</i> .....	107
<b>REPORTING AND ACTIONS .....</b>	<b>108</b>
REPORTING OVERVIEW .....	109
<i>Adding a New AOI</i> .....	110
<i>Summary</i> .....	110
<i>Coverage</i> .....	111
<i>Calculated Field Results</i> .....	112
<i>Actions</i> .....	113
SAVED VIEWS .....	114
<i>Defining a Work Package</i> .....	115
ACTIONS .....	116
<i>Filtering</i> .....	117
<i>Sorting</i> .....	117
<i>Scheduling an Action</i> .....	117
<i>Work Packages</i> .....	119
REPORT VIEWER .....	125
CREATING A REPORT .....	125
TYPES OF REPORT .....	128
<i>Calculated Fields Report</i> .....	128
<i>Containers Report</i> .....	130
<i>About the Key</i> .....	131
<i>Using Field Heat</i> .....	131
<i>Duplicates Reports</i> .....	133
<i>File Extensions Report</i> .....	138



<i>Files By Created/Last Accessed/Last Modified Reports</i> .....	139
<i>Files By Owner Report</i> .....	140
WORKING WITH REPORTS.....	142
<i>Chart tab</i> .....	143
<i>Data Table tab</i> .....	144
<i>Container List tab</i> .....	145
<i>File List tab</i> .....	146
<i>File Metadata Preview</i> .....	148
ACTIONS .....	151
<i>Export Data Table</i> .....	153
<i>Export Container List</i> .....	153
<i>Export File List</i> .....	153
<i>Delete</i> .....	155
<i>Quarantine</i> .....	155
<i>Migrate</i> .....	156
<i>Update Metadata</i> .....	160
<i>Markup</i> .....	160
<i>MIP Sensitivity Label (In Place)</i> .....	161
<i>File Extension Changes Following MIP Sensitivity Label Action</i> .....	162
<i>Preserving File Metadata Following MIP Sensitivity Label Action</i> .....	162
<i>Preserving SharePoint Metadata</i> .....	163
CUSTOM QUERIES.....	164
<i>Custom Query format</i> .....	164
<i>Executing Custom Queries</i> .....	165
MAPPING RULES.....	166
<i>Metadata Mappings</i> .....	166
CHARACTER MAPPINGS.....	169
<i>Adding a Character Mapping Set</i> .....	171
<i>Editing Mapping Actions</i> .....	171
<i>Deleting Mapping Actions</i> .....	171
<i>Editing a Character Mapping Set</i> .....	171
<i>Deleting Character Mapping Sets</i> .....	172
REPORTING SETTINGS .....	173
<i>Reporting Database Processing</i> .....	174
<i>Management Reporting Database Processing</i> .....	175
<i>Reporting Settings</i> .....	176
<i>Quarantine Locations</i> .....	176
<b>APPENDIX 1: INDEXING AND ANALYSIS.....</b>	<b>177</b>
OVERVIEW.....	177
SKIM PROCESSES.....	178
ANALYSIS PROCESSES.....	178
<i>File Retrieval</i> .....	178
<i>Duplicate Analysis</i> .....	179
<i>File Identification</i> .....	179
<i>File Contents Analysis</i> .....	179
<i>Other Analysis Types</i> .....	179
INDEX POST-PROCESSING .....	180
<b>APPENDIX 2: METADATA MODEL.....</b>	<b>181</b>



- APPENDIX 3: PREPARING FOR SHAREPOINT MIGRATION ..... 182**
  - PREPARING SHAREPOINT ..... 182
  - PREPARING CONTENT ..... 183
  - FIELD TYPES SUPPORTED FOR MIGRATION ..... 184
  - FIELD TYPES NOT SUPPORTED FOR MIGRATION ..... 184
- APPENDIX 4: FILES SUPPORTED FOR MIP SENSITIVITY LABELING ..... 185**
  - FILE FORMAT & EXTENSION RESTRICTIONS ..... 185
  - OTHER LABELING RESTRICTIONS..... 185
- APPENDIX 5: OPTIMIZING CONFIGURATION OF KEY MICROSOFT COMPONENTS..... 186**
  - ENABLE IIS DYNAMIC DATA COMPRESSION TO REDUCE IMPACT OF DATA EXPORT ON NETWORK ..... 186
  - CONFIGURE SQL SERVER FOR OPTIMAL PERFORMANCE ..... 187
  - CONFIGURE DATABASE VOLUME MAINTENANCE RIGHTS ..... 188
- APPENDIX 6: PRESERVING NTFS FILE OWNER..... 189**
  - MIPLABELACTIONPRESERVEFILEOWNER..... 189
  - MIPLABELACTIONERRORONFILEOWNERUPDATEFAILURE ..... 189
- APPENDIX 7: MANAGEMENT REPORTING ..... 190**
  - OVERVIEW..... 190
  - MRD DESIGN ..... 191
  - DISCOVERY CENTER CONFIGURATION..... 191
    - Calculated Fields* ..... 191
    - Storage Tiers* ..... 191
    - Repository Server Geographic Location* ..... 191
  - REPORTING TOOL DATA CONFIGURATION ..... 192
    - PowerPivot Essential Set Up* ..... 192
    - PowerPivot Optional Set Up* ..... 192
  - STARTING REPORT DESIGN ..... 192
    - Reporting on Data State* ..... 192
    - Reporting on Activity History* ..... 193
    - Working with Dimensions* ..... 193
    - Working with the Time Dimension* ..... 193
  - OLDEST COMPATIBLE MRD VERSIONS MATRIX ..... 194
- APPENDIX 8: CONNECTOR COMPATIBILITY SUMMARY ..... 195**
- APPENDIX 9: GLOSSARY ..... 198**
  - Activity* ..... 198
  - AN Administrator* ..... 198
  - Analysis* ..... 198
  - Analysis Queue* ..... 198
  - Analysis Schedule* ..... 198
  - Area of Interest (AOI)* ..... 198
  - Basic Metadata* ..... 198
  - Calculated Fields* ..... 199
  - Central Server* ..... 199
  - Chart* ..... 199
  - Classification* ..... 199
  - Cluster* ..... 199



<i>Container</i> .....	199
<i>Content Duplicate</i> .....	199
<i>Content Regular Expression Matches</i> .....	199
<i>Coverage</i> .....	199
<i>Crawl</i> .....	200
<i>Discovery Center</i> .....	200
<i>Diversity (a type of field score)</i> .....	200
<i>Duplicate Files</i> .....	200
<i>Excluded Location</i> .....	200
<i>Explicit Metadata</i> .....	200
<i>Extracted Data</i> .....	200
<i>Facet</i> .....	200
<i>Feature Pack</i> .....	200
<i>Field score</i> .....	201
<i>Field Source</i> .....	201
<i>File Extension</i> .....	201
<i>File Format</i> .....	201
<i>File Path Regular Expression Matches</i> .....	201
<i>File System Properties</i> .....	201
<i>First Value Field</i> .....	201
<i>Filter</i> .....	201
<i>Folder</i> .....	201
<i>Group</i> .....	202
<i>Heat</i> .....	202
<i>Index</i> .....	202
<i>Index Configuration</i> .....	202
<i>Index Ignored Location</i> .....	202
<i>Implicit Metadata</i> .....	202
<i>Index Security</i> .....	202
<i>Information Manager</i> .....	202
<i>Intensity</i> .....	202
<i>Keyword</i> .....	203
<i>Location</i> .....	203
<i>Management Reporting Database</i> .....	203
<i>Matched File Count</i> .....	203
<i>Metadata</i> .....	203
<i>Network Credentials</i> .....	203
<i>Node</i> .....	203
<i>Node Rule</i> .....	203
<i>Orphaned file</i> .....	203
<i>* Other</i> .....	204
<i>Remote Server</i> .....	204
<i>Report</i> .....	204
<i>Reporting Database</i> .....	204
<i>Resource Rank</i> .....	204
<i>Reviewer</i> .....	204
<i>Roles</i> .....	204
<i>ROT</i> .....	204
<i>Sample Files</i> .....	204
<i>Saved View</i> .....	205



<i>Similar Files</i> .....	205
<i>Skim</i> .....	205
<i>Solution</i> .....	205
<i>Storage Tier</i> .....	205
<i>Summary</i> .....	205
<i>Thematic Metadata</i> .....	205
<i>Title</i> .....	205
<i>Themes</i> .....	205
<i>Unclassified theme</i> .....	206
<i>User</i> .....	206

## Figures

FIGURE 1	ACTIVENAV METHODOLOGY .....	11
FIGURE 2	SIMPLE ACTIVENAV ARCHITECTURE .....	13
FIGURE 3	APPROACHES TO REVIEWING .....	18
FIGURE 4	SYSTEM OVERVIEW .....	19
FIGURE 5	THE HOME PAGE .....	21
FIGURE 6	HOME PAGE CHARTS.....	23
FIGURE 7	EXAMPLE WEB NOTIFICATIONS.....	25
FIGURE 8	APPLICATION LOGS.....	26
FIGURE 9	THE NETWORK MAP PAGE.....	28
FIGURE 10	ADDING A NETWORK LOCATION .....	30
FIGURE 11	EXCLUDING A LOCATION.....	31
FIGURE 12	STORAGE TIER TAB.....	34
FIGURE 13	SYSTEM SETTINGS – LICENSING .....	36
FIGURE 14	SYSTEM SETTINGS – USERS AND ROLES – ROLE MAPPING TAB .....	46
FIGURE 15	ADD ROLE MAPPING.....	46
FIGURE 16	SYSTEM SETTINGS – USERS AND ROLES – USER MANAGEMENT TAB .....	47
FIGURE 17	SYSTEM SETTINGS – CREDENTIAL MANAGEMENT.....	49
FIGURE 18	SYSTEM SETTINGS – DISCOVERY CENTER .....	51
FIGURE 19	ADD SCHEDULE CONSTRAINT DIALOG BOX .....	52
FIGURE 20	SYSTEM SETTINGS PAGE – EMAIL CONFIGURATION TAB.....	56
FIGURE 21	METADATA PAGE – CALCULATED FIELDS TAB .....	57
FIGURE 22	EXAMPLE OF PROXIMITY MATCHING RULES CALCULATED FIELD.....	63
FIGURE 23	METADATA - EXTRACTION RULES TAB.....	64
FIGURE 24	ADD EXTRACTION RULE DIALOG BOX .....	65
FIGURE 25	ADDING A BATCH OF KEYWORDS .....	68
FIGURE 26	EXAMPLE OF A KEYWORD-BASED EXTRACTION RULE .....	69
FIGURE 27	EXAMPLE OF A FILE PATH PATTERN MATCH EXTRACTION RULE .....	71
FIGURE 28	EXAMPLE OF A CONTENT PATTERN MATCH EXTRACTION RULE WITH A MASKING STRATEGY .....	71
FIGURE 29	EXAMPLE OF A REPOSITORY PROPERTY EXTRACTION RULE.....	72
FIGURE 30	EXAMPLE OF AN EMBEDDED WINDOWS FILE PROPERTY EXTRACTION RULE .....	73
FIGURE 31	THE BUILT IN SHAREPOINT FILE TYPE PROPERTY USED IN A REPOSITORY PROPERTY RULE .....	74
FIGURE 32	METADATA - CLASSIFICATIONS TAB .....	75
FIGURE 33	METADATA - METADATA VALUES IMPORT TAB.....	77
FIGURE 34	THE INDEXES PAGE – INDEX OVERVIEW TAB.....	79
FIGURE 35	EXAMPLE OF AN INDEX STATUS REPORT.....	83





FIGURE 36	INDEX CONFIGURATION TAB .....	87
FIGURE 37	INDEX CONFIGURATION – CALCULATED FIELDS TAB .....	89
FIGURE 38	INDEX CONFIGURATION – ANALYSIS FILTERS TAB .....	90
FIGURE 39	INDEX CONFIGURATION – DISCOVERY OPTIONS TAB .....	92
FIGURE 40	INDEX CONFIGURATION – ANALYSIS OPTIONS TAB .....	93
FIGURE 41	THE ADD INDEX DIALOG BOX .....	94
FIGURE 42	THE ADD MULTIPLE INDEXES DIALOG BOX .....	97
FIGURE 43	QUEUED TASKS ON THE CURRENT ACTIVITY PAGE .....	101
FIGURE 44	ACTIVITY HISTORY PAGE .....	103
FIGURE 45	AN INDEX PROCESSING TASK STATUS REPORT .....	105
FIGURE 46	AN ‘ACTIONS’ TASK STATUS REPORT .....	106
FIGURE 47	AN ‘MIP SENSITIVITY FILE LABELING’ STATUS REPORT .....	107
FIGURE 48	THE REPORTING OVERVIEW TAB .....	109
FIGURE 49	ADDING AN AREA OF INTEREST .....	110
FIGURE 50	SUMMARY TAB .....	111
FIGURE 51	COVERAGE TAB .....	112
FIGURE 52	CALCULATED FIELD RESULTS TAB .....	112
FIGURE 53	ACTIONS TAB .....	113
FIGURE 54	SAVED VIEWS TAB .....	114
FIGURE 55	CREATE A WORK PACKAGE .....	116
FIGURE 56	ACTIONS TAB .....	118
FIGURE 57	WORK PACKAGES INFORMATION MANAGER VIEW .....	119
FIGURE 58	WORK PACKAGES REVIEWER VIEW .....	119
FIGURE 59	ACTIVATE WORK PACKAGE VIA ACTIVATE ICON .....	121
FIGURE 60	ACTIVATE WORK PACKAGE VIA STATUS .....	121
FIGURE 61	DEACTIVATING A WORK PACKAGE .....	122
FIGURE 62	DEACTIVATING A WORK PACKAGE VIA REVIEW PACKAGE STATUS .....	122
FIGURE 63	DEACTIVATING A WORK PACKAGE VIA REPORT VIEWER .....	123
FIGURE 64	WORK PACKAGES TAB .....	123
FIGURE 65	RESETTING A WORK PACKAGE .....	124
FIGURE 66	DELETING A WORK PACKAGE .....	124
FIGURE 67	WORK PACKAGES TAB .....	125
FIGURE 68	DEFINE VIEW DIALOG BOX .....	127
FIGURE 69	CALCULATED FIELDS REPORT .....	129
FIGURE 70	CONTAINERS REPORT (WITHOUT A FIELD HEAT OVERLAY) .....	130
FIGURE 71	CONTAINERS REPORT WITH FIELD HEAT OVERLAY .....	132
FIGURE 72	USING THE AREA OF INTEREST MASTER SELECTION STRATEGY .....	134
FIGURE 73	EXAMPLE OF A FILE DUPLICATES REPORT .....	136
FIGURE 74	EXAMPLE OF A CONTENT AND FILE DUPLICATES REPORT .....	137
FIGURE 75	FILE EXTENSIONS REPORT .....	138
FIGURE 76	FILES BY CREATED/LAST ACCESSED/LAST MODIFIED REPORT .....	140
FIGURE 77	FILES BY OWNER REPORT .....	141
FIGURE 78	THE REPORT VIEWER/CHART TAB .....	143
FIGURE 79	DATA TABLE TAB .....	145
FIGURE 80	CONTAINER LIST TAB .....	146
FIGURE 81	EXAMPLE FILE LIST .....	147
FIGURE 82	FILE METADATA PREVIEW: BASIC METADATA TAB .....	149
FIGURE 83	THEMES & SUMMARIES TAB .....	150
FIGURE 84	CALCULATED FIELDS TAB .....	150
FIGURE 85	MARKUP TAB .....	151



FIGURE 86	THE ACTIONS MENU.....	152
FIGURE 87	EXPORT FOLDERS ACTION .....	153
FIGURE 88	SELECT FILE EXPORT COLUMNS .....	154
FIGURE 89	MIGRATE ACTION.....	159
FIGURE 90	MARKUP ACTION .....	160
FIGURE 91	MIP SENSITIVITY LABEL ACTION .....	162
FIGURE 92	CUSTOM QUERIES PAGE.....	165
FIGURE 93	REPORTING AND ACTIONS PAGE – MAPPING RULES – METADATA MAPPINGS.....	166
FIGURE 94	BASIC METADATA MAPPINGS.....	167
FIGURE 95	MAPPING OPTIONS.....	168
FIGURE 96	REPORTING AND ACTIONS PAGE – MAPPING RULES – CHARACTER MAPPINGS .....	169
FIGURE 97	THE SHAREPOINT SPECIAL CHARACTER REPLACEMENTS MAPPING SET.....	172
FIGURE 98	REPORTING AND ACTIONS PAGE – REPORTING SETTINGS.....	173
FIGURE 99	INDEX PROCESSING STEPS .....	177
FIGURE 100	ANALYSIS PROCESSING FLOW CHART .....	178
FIGURE 101	METADATA MODEL .....	181
FIGURE 102	INSTALLING STATIC CONTENT COMPRESSION .....	186
FIGURE 103	CONFIGURING COMPRESSION OPTIONS.....	187
FIGURE 104	CONFIGURING DATABASE VOLUME MANAGEMENT RIGHTS .....	188
FIGURE 105	MANAGEMENT REPORTING DATABASE ARCHITECTURE .....	190

**Tables**

TABLE 1	NOTIFICATIONS RECEIVED BY EACH OF THE AN ROLES.....	24
TABLE 2	APPLICATION LOGS AND THEIR USES.....	27
TABLE 3	NETWORK MAP STATUS CODES AND DESCRIPTIONS .....	33
TABLE 4	SOLUTIONS AND FEATURE PACKS .....	37
TABLE 5	LICENSE PACKS.....	39
TABLE 6	SYSTEM USAGE INFORMATION .....	43
TABLE 7	PERMISSION REQUIRED FOR VARIOUS TASKS .....	44
TABLE 8	SYSTEM SETTINGS – USER ACCESS .....	48
TABLE 9	CLASSIFICATION FIELDS: VALUES AND PATH SETTINGS.....	60
TABLE 10	UNDERSTANDING MATCHING STRATEGIES AND CALCULATED FIELD TYPES .....	62
TABLE 11	RETRIEVAL MESSAGES.....	84
TABLE 12	ANALYSIS MESSAGES .....	85
TABLE 13	OPTIMIZING AN INDEX.....	97
TABLE 14	HOW MIGRATION SETTINGS AFFECT FOLDER PERMISSIONS .....	157
TABLE 15	MIP SENSITIVITY LABEL METADATA PRESERVATION .....	163
TABLE 16	MAPPING OF BASIC FILE PROPERTIES FOLLOWING MIGRATION.....	167
TABLE 17	STANDARD CHARACTER MAPPINGS SETS.....	170
TABLE 18	SHAREPOINT FIELD TYPES SUPPORTED FOR MIGRATION.....	184
TABLE 19	SHAREPOINT FIELD TYPES NOT SUPPORTED FOR MIGRATION .....	184
TABLE 20	SCOPE OF FACTS TABLES.....	193
TABLE 21	OLDEST COMPATIBLE MRD VERSIONS .....	194



# Introduction

Our Discovery Center file analysis platform provides organizations with the ability to rapidly discover and understand information compliance and quality problems in chaotic unstructured information stores. Discovery Center analyzes electronic files in place to create an efficient index containing key metadata for use in a wide range of information governance scenarios from content clean-up and disposal, through compliance and migration to continual content governance. It works with a range of modules and rules packs which are combined to enable specific solutions for each customer. The entire product is designed around a structured methodology which enables clients to proactively address their legacy information as part of an ongoing program of governance.

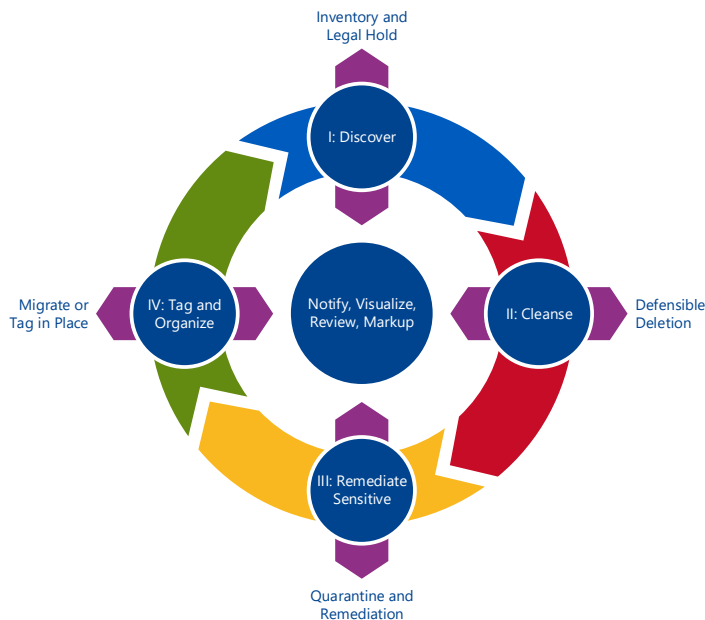


Figure 1 ActiveNav Methodology

**Discover.** Discovery and reporting delivers a clear understanding of what information currently exists: volumes of data, its breakdown, key topics and any ROT. Discovery results can be explored, drilled into and exported to ensure you know exactly what information you have.

**Cleanse.** Actions remove ROT in bulk, based on default rule sets and customizable policies. This includes analytics and tools to examine and remediate duplicates both across and between repositories. In file shares, more than 30% of content is often removed at this stage.

**Remediate Sensitive.** Files containing leaked PII, protective markings and other sensitive data are identified and classified across discovered content. Structured review workflows, markup, actions and audit trails provide a defensible path to their remediation.

**Tag and Organize.** Legacy content is often 'chaotic' with poor organization and little or no metadata. Discovery Center extracts new, consistent metadata and can define more meaningful filing structures. The results can be reviewed and refined before they are mapped to a new repository for direct migration or via a third-party tool.



# Product Overview

## Modules and Components

To allow tailoring to specific solutions, Discovery Center is modularized; modules are delivered in a single executable installation file and enabled by license configuration:



Discovery Center



Analysis



Delete and Quarantine



Tag and Organize



Connectors

Discovery Center provides capabilities from system management through metadata management to indexing and reporting. Analysis extends the indexing capabilities of Discovery Center, adding analytics of file contents whilst the actions enable the user to Delete and Quarantine or Tag and Organize files in bulk. Finally, connectors enable Discovery Center to work across a range of different information repositories.

Discovery Center is supported by the Discovery Center Workbench client application for the design and modelling of classifications. Workbench is usually installed for a small number of specialist users.



# Technologies

Discovery Center is built upon Microsoft technologies, exploiting the capabilities of Windows Server, Internet Information Services (IIS) and SQL Server to deliver a high performance and scalable solution that can be readily deployed and integrated into any enterprise environment. Windows Server provides the IIS web application host operating system and drives all discovery and analysis capabilities through the Discovery Center web application. SQL Server supports the Discovery Center OLTP database for the scalable storage of system configuration settings and index results. SQL Server Analysis Services deploys an OLAP cube for high performance reporting which provides web-based charting for visualization of discovery and analysis results. Discovery Center is developed using .Net Framework 4.7 with C# and JavaScript.

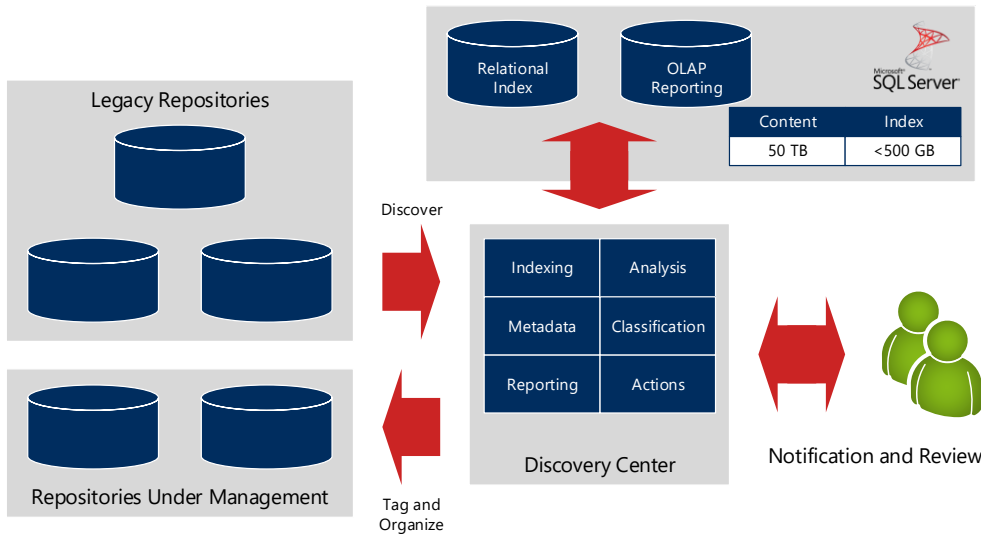


Figure 2 Simple ActiveNav Architecture



# System Performance

ActiveNav Discovery Center is designed to index, analyze and act upon large numbers and volumes of files across any organization’s network infrastructure. Like all similar technologies, performance is largely governed by the performance of the underlying network and the specification of host hardware in use.

Scenarios for well configured systems and quality networks can achieve file discovery rates of between 500,000 and 1,000,000 files per hour (or 1TB per 24 hr period).



# Concepts and Principles

The following general concept and principles are used throughout ActiveNav's software. A clear understanding of each will improve the software's effectiveness and user efficiency.

## Security

### Roles and Responsibilities

Roles provide users with access to different levels of capability according to their Windows credentials configured by the System Administrator. It is recommended that specific Windows groups be defined for these roles and, where appropriate, to allow the allocation of credentials to indexes.

An individual or group can be allocated one or more roles; Discovery Center provides four different roles as follows:

- **System Administrator**  
The System Administrator configures and manages the way the Discovery Center interacts with the host environment. This includes defining security groups, managing schedule constraints and controlling the network map to prevent unwanted indexing.
- **AN Administrator**  
The AN Administrator defines and schedules indexes for analysis and manages the extraction rules and calculated fields needed to extract attributes from analyzed files.
- **Information Manager**  
An Information Manager can use the results of Discovery Center indexing and analysis processes to generate reports and act on them according to information policies. An Information Manager may define Areas of Interest, view reports according to index permissions and act on those reports.
- **Reviewers**  
Reviewers are provided access to reporting only to support the review of analysis results and subsequent action by subject matter experts.

### Access to Information

ActiveNav's Discovery Center is designed to discover and aggregate large volumes of information to enable authorized users to directly address legacy information and information quality problems across a range of different environments and sources. Care should be taken to ensure that Discovery Center users are allocated to the correct roles and that services (such as the scheduler) and indexes are provided credentials as described in the Installation Guide.

All Discovery Center services and activities are controlled using Windows authentication.

### Privileged User

In order to enable reporting and actions across large amounts of information and to remove constraints inherent in many existing or legacy access control lists, Discovery Center uses the principle of the 'privileged user' such that a small group of authorized users are provided broad rights to act. Discovery Center users therefore operate within the following privileged security framework:

- **Access to Functionality**  
Users are provided access to Discovery Center functionality through the four specific user access groups described above.



- **Access to Source Files**  
The Discovery Center uses credentials assigned to network map locations in order to discover and analyze files. If these credentials are not provided, the Discovery Center will use the credentials provided to the ActiveNav scheduler service as detailed in the Installation Guide.
- **Access to Analysis Results**  
Users are granted access to analysis results provided they have the necessary credentials assigned to the corresponding network location(s).
- **Acting on Files**  
To move or delete files, the Discovery Center uses the credentials provided to the corresponding network location, or the scheduler service as detailed in the Installation Guide.

## Network Map, Indexes and Areas of Interest

ActiveNav's Discovery Center makes use of three features through which to discover, analyze and present information value. Each of these features interacts as follows:

- **Network Map**  
The network map provides a representation of the topography and structure of discovered information sources and is the primary means by which users explore that information. It also enables AN Administrators to plan index configuration based upon file distribution across servers and containers. Access to areas of the network map is controlled by the System Administrator by assigning credentials to network map locations. If these credentials are not provided, Discovery Center uses the credentials provided to the ActiveNav scheduler service as detailed in the Installation Guide.
- **Indexes**  
An Index defines, starting from a single location in the network map, the information that will be analyzed. An *Index Configuration* is defined to control how information is analyzed and what metadata attributes are collected. An index 'skim' collects only basic file properties (such as file extensions and modified dates) and is completed very rapidly. In order to collect additional attributes, such as themes or metadata in calculated fields, indexes must be configured to read file contents. This process, known as analysis, retrieves files from the source and therefore proceeds much more slowly than a skim. Attributes gathered by an index are stored for use in reporting, classification and migration. Indexes must be configured so that they do not overlap.
- **Areas of Interest (AOI)**  
Once a skim or analysis has been completed, the network map may be explored to report upon and discover file attributes from a given container. In order to explore results combined from more than one container, an AOI must be defined. AOIs allow information to be explored based upon concepts that are relevant to the organization rather than actual information locations. For example, where a specific business function has files stored in two different shares, an AOI can be created to report upon that information a whole.

## Skims, Analysis and Classification

The Discovery Center indexes control the analysis of information from discovered file locations such as a file share, a specific folder on a file server or (with the appropriate connector) a SharePoint document library. When an index is run, its configuration determines which analysis tasks are to be completed, from basic file property collection to duplication or thematic analysis. The speed with which an index is completed depends upon a number of factors including:

- Number of files within the index.
- Performance of the network connection between the file source and the Discovery Center.
- Number and complexity of analysis tasks to be completed by the index.
- Hardware and virtualization specification for the Discovery Center and database hosts.



Indexes run in up to three phases: "skim, analyze, classify" depending upon their configuration. First the index discovers the topography of the source information, reading container and file attributes without opening the files themselves. This initial process is known as a skim and collects the following basic properties:

- File path, folders and file name.
- File extension, date created, date modified and file size.

Once a skim has been completed, additional index settings will cause analysis to take place; note that analysis tasks temporarily retrieve files from the source to the Discovery Center cache so that they can be efficiently analyzed.

- Duplication and similarity analysis.
- File format analysis.
- Thematic analysis.
- Calculated fields.

Once all analysis tasks have been completed, classification takes place, populating all classification calculated fields using the results from the index process.

## Calculated Fields and Extraction Rules

The Discovery Center collects information about files in an index and stores that information in its SQL database as metadata. Basic file property metadata is collected during a skim whilst other types of metadata are either extracted from a file's contents (such as themes or summaries) or calculated based upon rules. Understanding how these rules interact to produce valuable metadata is key to getting more advanced results from an ActiveNav project.

Extraction rules are the basic building block for extracting and deriving custom metadata from a file's contents. Extraction rules define how Discovery Center's analyzers work with text or other properties of a file and different extraction rules can be combined to increase the reliability of metadata creation. Discovery Center provides four different types of extraction rule for this purpose:

- File property rules read values from embedded Microsoft Office file properties in each document, for example, *Created Date, Author, Subject* or EXIF values from image files such as camera settings, copyright and date/time information.
- File path regular expressions look for pre-defined text patterns appearing anywhere in the file path and extract the first instance of matching text from the file path. This technique may be used to match certain folder names or look for metadata embedded by the user within a file name.
- File content regular expressions seek patterns within the text of analyzed files and extract the first instance of matching text from the file contents. This technique can be used to find, for example, document or process reference numbers from file text.
- Keyword matching searches for specified keywords (including synonyms) within file content.
- Repository property matches metadata fields from a source repository such as document library columns in SharePoint. Repository properties only function when a suitable connector has been configured for the source in question.

Calculated fields use a classification structure or a set of extraction rules to derive a value (or values) useful for reporting or as metadata to support migration. Neither classification structures nor extraction rules can be used to create metadata without first being applied as part of a calculated field:

- Matching rules are used to capture the results of successful extraction rules. Each file in an index has metadata assigned according to either the value(s) of the first successful extraction rule or all listed extraction rules.
- Classification fields are used to assign the names of classification nodes as metadata values. Any file successfully classified will have one or more values assigned dependent upon the classification node rules in use.





# Microsoft Information Protection Sensitivity Labels

Discovery Center provides the ability to read and apply Microsoft Information Protection (MIP) Sensitivity Labels for files under management. These features require a Microsoft 365 tenant with an E3 subscription, the creation of an Azure Active Directory Enterprise Application and an Azure Active Directory User Account. For more details on how to set up this integration see **Appendix 10** of the **Discovery Center Installation Guide**.

Once the **System Settings > Discovery Center > MIP Settings** have been applied, Discovery Center can be configured to read Sensitivity Labels from files during a Textual Analysis index. The first step is to create an Extraction Rule:

- **Type:** Embedded file property
- **Data Type:** String
- **Property Types:** MIP Label Property
- **Property:** MIP Sensitivity Label

The next step is to create a Calculated Field based on the new Extraction Rule, ensuring that the value is available in Reports. Once the new Calculated Field is added to an Index Configuration, a new Index can be created. During the Analysis phase of indexing, Discovery Center will use the User Account and other configured MIP Integration Settings that are defined in the Discovery Center System Settings page to connect to the Enterprise Application and query the MIP service for any Sensitivity Label applied to supported file types. If a Sensitivity Label is reported for a file, the Calculated Field will be populated with its value which will subsequently be available for reporting.

Applying MIP Sensitivity Labels to files with Discovery Center is accomplished either during a migration action, or in place using an MIP Sensitivity Label action, by assigning a Calculated Field as the source for the MIP Sensitivity Label. If the value of the Calculated Field matches the name of a Sensitivity Label defined in the M365 tenant and the target files are supported by MIP, the action will apply the Sensitivity Label to the files in the target repository. The Calculated Field to use for this mapping is chosen at the point of configuring the action and only Calculated Fields that hold a single value are available for selection, as files may only have a single MIP Sensitivity Label applied.

Only certain file formats are valid for applying MIP Sensitivity Labels, with some of those formats only valid for labeling that applies RMS encryption to the file. For more information on supported file formats, see Appendix 4.

## Reviewing

The Information Manager is responsible for interrogating the system to generate reports, for example: to identify policy violations, redundant or duplicate files, or prepare for migrating or archiving files. Generally, the Information Manager needs to work closely with the business users or file owners responsible for the network locations under investigation. Discovery Center provides a *Reviewer* user role, allowing business users limited access to view and mark up reports created by the Information Manager. A review process can be led by the Information Manager or by the Reviewer (the business user) and Discovery Center provides tools for each of these approaches (see Figure 3).

### Using Work Packages

This method offers a fully-audited, structured approach to reviewing. The Information Manager assigns a **Work Package** to a Reviewer whenever input is required. A Work Package is created from a Saved View and has a description guiding the Reviewer about the information and decisions needed, for example: identifying files to be deleted. Each Work Package has a deadline for the work to be completed. Notifications inform the Information Manager and Reviewer about the status of the Work Package from its inception to completion and approval.



## Sharing Saved Views

This method offers a more flexible approach to reviewing. The Information Manager carries out an analysis of a network area and saves a report view, allowing access to the Reviewer(s) responsible for the network location. A Reviewer uses markup fields to identify actions for the IM to implement. Notifications inform the Information Manager and Reviewer about changes to the saved View and the application of any actions such as markup.

When assigning Work Packages or sharing Saved Views the Information Manager has the option to provide any local or Active Directory Windows group that have been assigned the Reviewer or Information Manager role in Discovery Center (See *Users and Roles* page 23). This has the effect of allowing all members of the given group access to Saved Views or Work Packages respectively, without the need to assign them all individually. In this case each member of the group will also receive individual notifications in regard to the Saved View or Work Package in question, dependent on their personal notification settings within Discovery Center (See *User Profile* page 24).

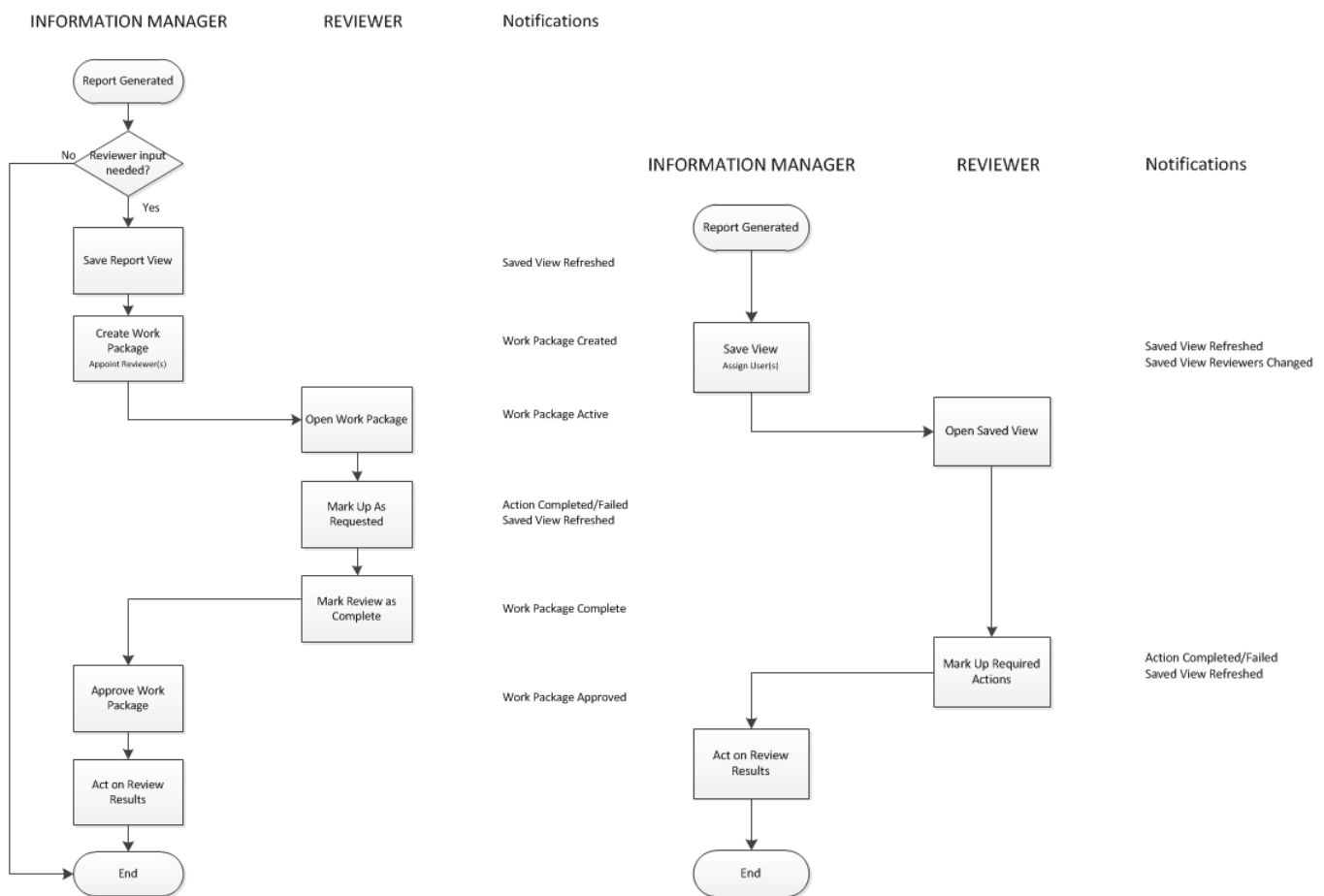


Figure 3 Approaches to reviewing



# System Overview

The ActiveNav suite consists of a range of component applications and supporting services provided by Microsoft Windows and SQL Server technologies. The functions of the main components are:

- **Discovery Center**  
A web application delivered through Microsoft Internet Information Services (IIS), through which the user interacts with and manages the indexing and analysis of files in source repositories and reporting for information cleansing, migration and governance.
- **Discovery Engine (Scheduler Service)**  
A Windows service (ActiveNav Scheduler) that controls the sequencing, execution and termination of Discovery Center tasks.
- **SQL Server Database**  
Indexes stored within a SQL Server database contain the results of all ActiveNav analyses; the database lies at the heart of ActiveNav and supports all functions and features.
- **SQL Server Analysis Services (SSAS)**  
SSAS provides a reporting database to support interactive charting for the Discovery Center.
- **Client Applications**  
ActiveNav client applications provide specialist capabilities for duplication cleansing and classification rules design and deployment.

The following diagram on shows the key relationships between these components:

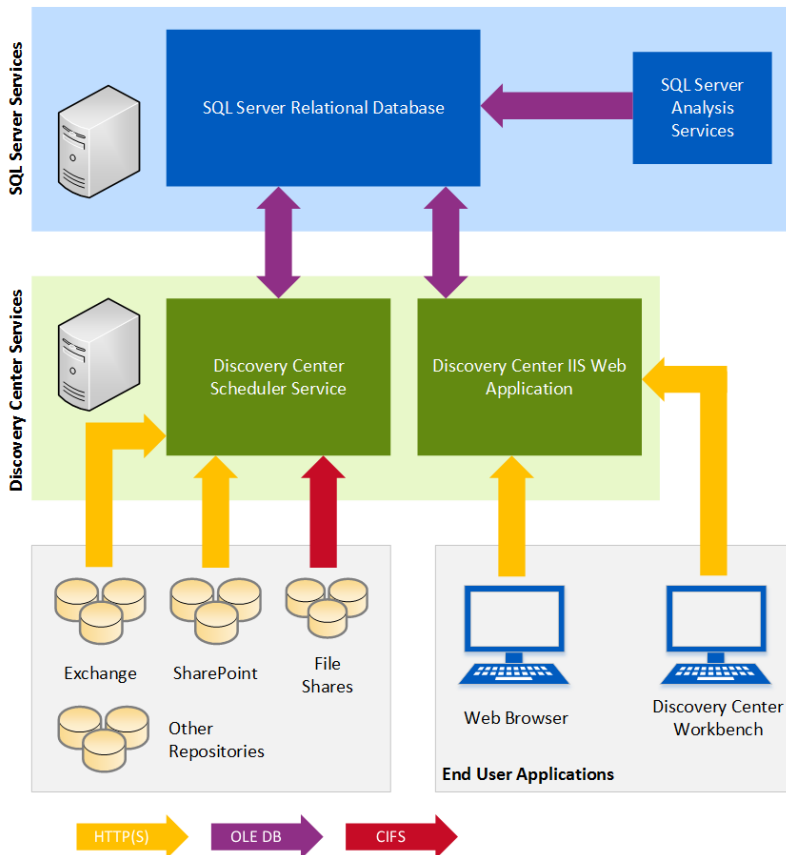


Figure 4 System Overview



# Discovery Center

The main functions of the Discovery Center are:

- **System Management**  
System management features include user role mapping, definition of system constraints, the establishment and maintenance of the network map and connector configuration.
- **Index Management**  
The Discovery Center provides full control over all indexing options (such as the selection of analysis types, allocation of index credentials and creation of index to metadata mappings) as well as index scheduling.
- **Metadata Management**  
Metadata fields and supporting metadata extraction rules are configured within the Discovery Center and then made available to indexes for population. Metadata fields and extraction rules are fully user definable.
- **Reporting**  
The Discovery Center provides comprehensive reports which allow Information Managers and Reviewers to explore the value of their electronic information and pinpoint information quality problems. Reporting is supported across the full range of available metadata and default reports provide for the analysis of file distribution by type, file ages, duplication and folder usage. Report views can be saved and recalled as necessary.

To start the Discovery Center select:

Start > All Programs > Active Navigation > Discovery Center

The Discovery Center's *Home* page is displayed. Depending on the role associated with your Windows account, some or all of the following tabbed pages may also be available:

- Network Map (System Administrators only)
- System Settings (System Administrators only)
- Metadata (AN Administrators only)
- Indexes (AN Administrators only)
- Current Activity (AN Administrators and System Administrators only)
- Reporting and Actions (AN Administrators, Information Managers and Reviewers)



# Home

The *Home* page provides links to frequently used functions related to your assigned Discovery Center roles.

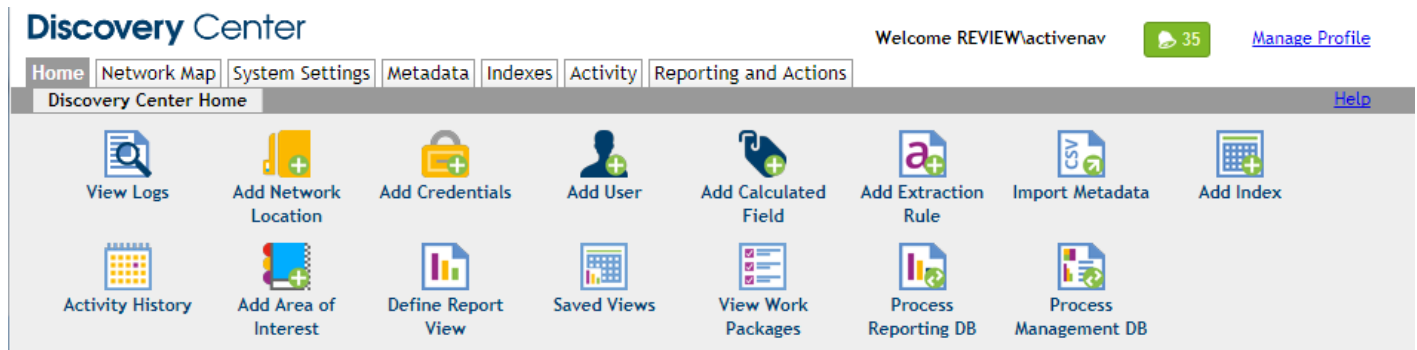


Figure 5 The Home page

## Discovery Center Home - System Administration

The System Administrator configures and manages the way the Discovery Center interacts with the host environment. This includes defining security groups, assigning users to Discovery Center roles, managing credentials, setting schedule constraints and controlling the network map to determine the network locations available for analysis and to prevent unwanted indexing.

In addition to the Home page, a System Administrator has access to the Network Map, System Settings and Activity pages.

Use the following links to administer the Network Map and user access:

- View Logs  
The application log files can help diagnose problems with the system (see page 26).
- Add Network Location (see page 30)
- Add Credentials (see page 49)
- Add User (see page 47)

## Discovery Center Home - AN Administration

The AN Administrator defines and schedules indexes for analysis and sets up the extraction rules and calculated fields needed to extract attributes from analyzed files. In addition to the *Home* page, an AN Administrator also has access to:

- The Metadata, Indexes and Activity pages
- The Reporting Settings and Migration Mappings tabs on the Reporting and Actions page.

Use the links to administer indexes and to configure calculated fields:

- View Logs  
The application log files can help diagnose problems with the system (see page 26).
- Add Calculated Field (see page 59)
- Add Extraction Rule (see page 65)
- Import Metadata (see page 76)



- [Add Index](#) (see page 87)
- [Process Reporting DB](#) (see page 174)
- [Process Management DB](#) (see page 175)

## Discovery Center Home - Information Management

The Information Manager is responsible for interrogating the system to generate reports, identify policy violations and remove redundant files. In addition to the Home page, an Information Manager has access to the Reporting and Actions page.

Use these links to prepare reports and to action your content:

- [Add Area of Interest](#) (see page 109)
- [Define Report View](#) (see page 125)
- [Saved Views](#) (see page 114)
- [View Work Packages](#) (see page 118)

## Discovery Center Home - Reviewing

A Reviewer is a business user or file owner responsible for one or more network locations. The Reviewer assists the Information Manager in the analysis of results in their area of responsibility (see *Reviewing*, page 16).

Use this link to explore results and tasks assigned to you:

- [Work Packages](#) (see page 115)
- [Saved Views](#) (see page 114)



## Dashboard Charts

The Home page also displays dashboard charts summarizing data from all indexed locations. The range of charts shown will depend on the features licensed for your installation and the status of your Reporting Database.

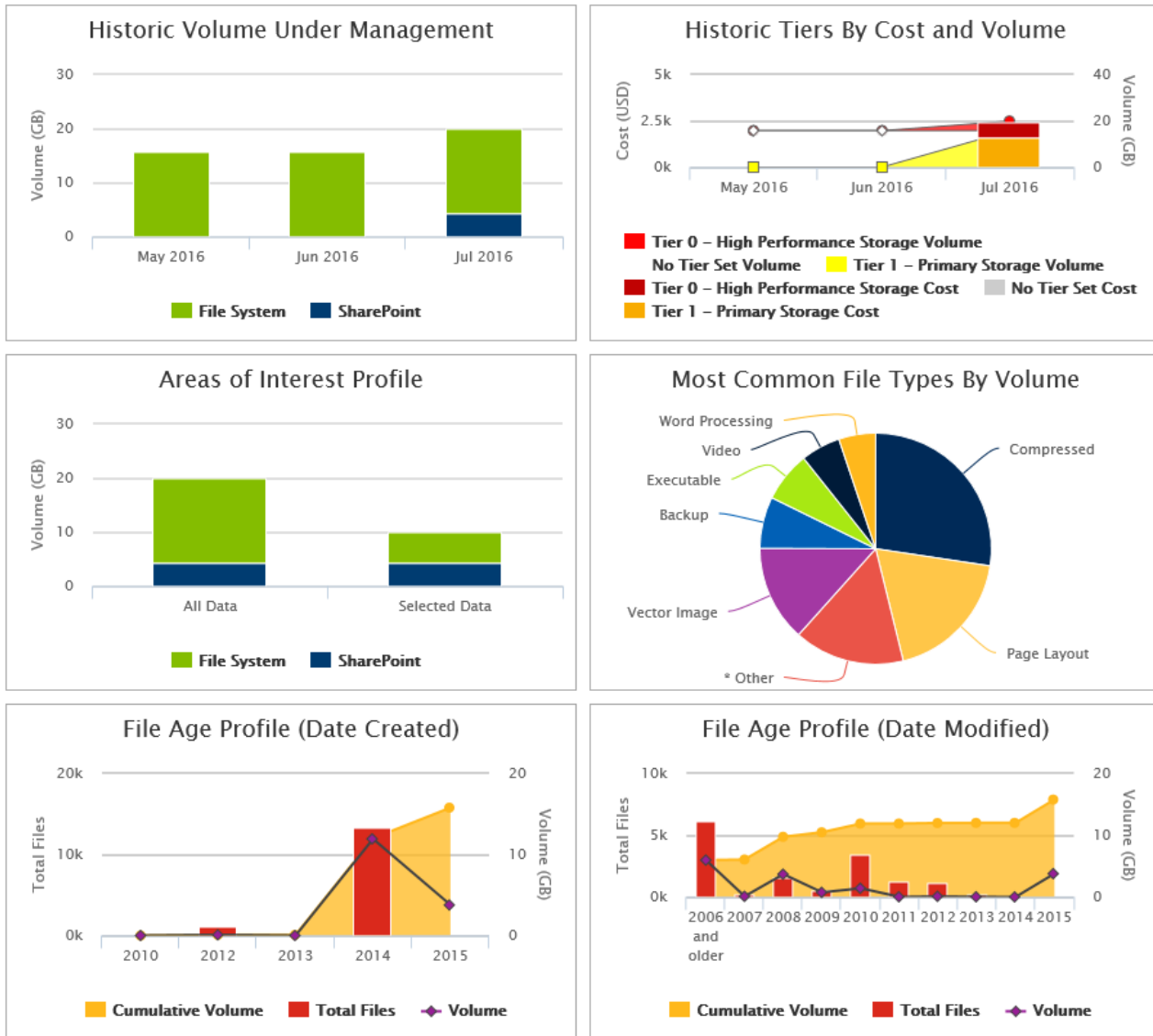


Figure 6 Home page charts

**Note.** The Historic Volume Under Management and Historic Tiers By Cost and Volume charts are only displayed if the Management Reporting license is active. The Management Reporting Database accumulates historical snapshots of the primary database state. In addition to the charts shown on the Home page, you can monitor information management metrics such as key calculated fields, storage costs, activities performed using third party business intelligence and reporting products (see page 189).



# User Profile

The User Profile page allows you to edit your contact details and manage notifications:

- User Details**  
 Edit your first and last names and contact email address. You can choose which system notifications, if any, are sent to the email address (see below).
- Notification Management**  
 Choose which web and email notifications you want to receive (see below).

## Notifications

The notifications you receive are determined by the Discovery Center profiles associated with your user account (see table below).

Table 1 Notifications received by each of the AN roles

Notification	AN Administrator	Information Manager	Reviewer	System Admin
Action Completed	✗	✓	✓	✗
Action Failed	✗	✓	✓	✗
Index Export Completed	✓	✗	✗	✗
Index Export Failed	✓	✗	✗	✗
Index Import Completed	✓	✗	✗	✗
Index Import Failed	✓	✗	✗	✗
Index Processing Completed	✓	✗	✗	✗
Index Processing Failed	✓	✗	✗	✗
Index Reclassification Completed	✓	✗	✗	✗
Index Reclassification Failed	✓	✗	✗	✗
Metadata Import Completed	✓	✗	✗	✗
Metadata Import Failed	✓	✗	✗	✗
Process Reporting Database Completed	✗	✓	✗	✗
Process Reporting Database Failed	✗	✓	✗	✗
Saved View Refreshed	✗	✓	✓	✗





Notification	AN Administrator	Information Manager	Reviewer	System Admin
Saved View Reviewers Changed	✗	✓	✓	✗
Work Package Created	✗	✓	✓	✗
Work Package Refreshed	✗	✓	✓	✗
Work Package Status Changed	✗	✓	✓	✗

Web notifications are displayed by clicking on the button in the top right of the window (see Figure 7). These are color-coded according to their importance: red indicates the failure of a process or action; green indicates the successful completion of a process or action, and blue is used to convey progress information about Work Packages. Events that complete with warnings are shown with an amber background. The notifications button takes the colour of the most significant notification in your personal list (in the sequence red > green > blue). Click on the **More...** link to investigate a notification. You will need to provide username and password authentication to proceed.

Your selections on the *Notification Management* tab determine which notifications you receive. By default, email and web notifications are enabled for error conditions and for saved view related updates. To receive email notifications, an email address must be set on the *User Details* tab.

Notifications are an important part of the review process, telling Information Managers and Reviewers about the progress and status of Work Packages.

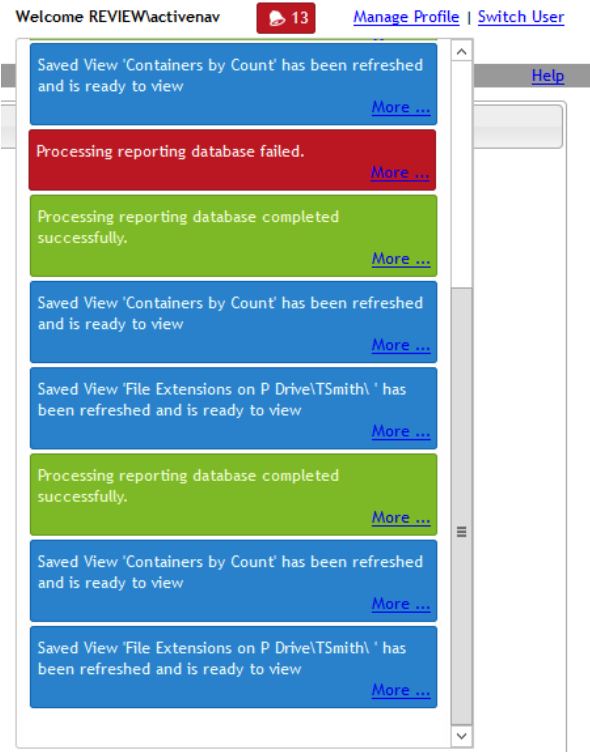


Figure 7 Example Web Notifications



# Application Logs

Should an error occur during the normal course of operation, immediate diagnosis and resolution can be sought by reviewing the relevant application logs. Application logs are accessible to the System Administrator role from the *Home* tab via the *View Logs* link as shown below. All Discovery Center application logs are stored in the Discovery Center installation folder.

## Applications Logs

[Web UI and Application Overview Log](#)

Detailed application logs:

Log file	File size	Last modified time	
Analysis.log	128.90 KB	12/15/2015 9:42:10 AM	<a href="#">Download</a>
Scheduler.log	4.94 MB	7/8/2016 8:47:57 AM	<a href="#">Download</a>
SchedulerTasks.log	74.26 KB	6/29/2016 2:49:43 PM	<a href="#">Download</a>
Skimmer.log	73.39 KB	12/15/2015 9:40:56 AM	<a href="#">Download</a>

[Web UI and Application Overview Log](#)

Figure 8 Application logs

The *Web UI and Application Overview Log* is stored in the Discovery Center database and is browsed directly in the Discovery Center browser interface. Other application logs are stored in text files which can be downloaded from the links shown above. Each type of log will be added to as new events arise until that log reaches 10MB in size, at which point a new log will be started named with a suitable date-time stamp. Old logs can be safely deleted from the logs folder if necessary.

In the normal course of operation, the Discovery Center logs a wide range of events:

- **General.** Any event not captured under other categories. The severity of the event will be recorded and indicates whether any action might be required to resolve a problem.
- **Milestone.** Information events included to record progress of a particular task; for indexing (skim and analysis), milestones show how the index progressed with time.
- **Performance.** Usually logged at the end of an analysis or action task, these entries summarize how quickly the task was completed and provide valuable information to help spot bottlenecks or expensive operations.

Event entries that indicate a potential problem or failure of part of a process have a severity assigned as follows:

- **Warning.** Some non-critical process failed but as a single failure, that task would not normally impair product functionality. Warning errors do not usually require attention unless the failing task is important to the project being undertaken.
- **Error.** A process failed which is likely to have a detrimental impact on product functionality. Error entries should usually be referred to ActiveNav support for triage and diagnosis.

The following table describes the logs for the Discovery Center and its supporting products. It also describes the product functions handled by each type of log.



**Table 2 Application logs and their uses**

Log Name	Uses and Product Functions	Storage Location
Web UI and Application Overview	Go here first for any Discovery Center application error. This log provides an overview of all processes and tasks.	Discovery Center database
Scheduler	Records the timing and sequence of skim, analysis, classification or action tasks. Go here to diagnose issues relating to the management of classification, actions, indexes and reporting database processing. Also go here for action performance statistics.	Log file location selected during installation – default is: \\Active Navigation\Discovery Center\Logs
Scheduler Tasks	Records issues relating to index import and export and Management Reporting database processing.	Log file location selected during installation – default is: \\Active Navigation\Discovery Center\Logs
Skimmer	Skimmer errors including network discovery.	Log file location selected during installation – default is: \\Active Navigation\Discovery Center\Logs
Analysis	Analysis progress and performance statistics as well as any analysis errors.	Log file location selected during installation – default is: \\Active Navigation\Discovery Center\Logs
Windows Application Log	Provides an understanding of issues concerning third party applications supporting ActiveNav (SQL Server and IIS) and windows-related problems such as user or SQL Server authentication.	Windows Event Viewer



# Network Map

The Network Map page has two tabs:

- **Network Map**  
View servers available for analysis, indexing, reporting and actions.
- **Storage Tiers**  
Assign storage costs to servers in Network Map (or shares for file system).

## Network Map

The Network Map lists servers available for analysis, indexing, reporting and actions.

The screenshot shows the Discovery Center interface. The top navigation bar includes 'Home', 'Network Map', 'System Settings', 'Metadata', 'Indexes', 'Activity', and 'Reporting and Actions'. The 'Network Map' tab is active. The main content area is divided into two panels. The left panel, titled 'Discovered Locations', shows a tree view of discovered locations. The right panel, titled 'Location Summary', provides details for the selected location.

**Discovered Locations**

- localhost [764.92 MB]
- \_temp [19.46 MB]
- additional test data [247.24 MB]
- AN Test Files [338.67 MB]
- DE1538 Long Filename [9.06 KB]
- DeepFolder [282.16 KB]
- Migration Destination [0 B]
- test new docs [184.02 KB]
- test old docs full [79.54 MB]
- test old docs [79.54 MB]
- cowfish.pond.private.activenavigation.com:80 [86.12 MB]
- AN Test Files [84.36 MB]
- ANTMP\_NewlyCreatedColumnLibrary [0 B]
- GJH Nasty Library [1.49 KB]
- MandatoryFieldLibrary [0 B]
- Metadata Test Library LP [20.39 KB]
- Metadata Test Library RAA [0 B]
- Metadata Test Library SJM [0 B]
- Metadata Test Library [1.03 MB]
- Metadata Site [0.77 KB]

**Location Summary**

Location:	\\localhost\
Location Type:	Server
Connector Name:	File System Connector
Retrieval Status:	The resource has been skimmed by Active Navigation.
Size:	764.92 MB
Number of Folders:	2,697
Number of Files:	13,322
Associated Items	
Index Start Location:	localhost everything
Local Credentials:	No local credentials have been set
Inherited Credentials:	No credentials set for parent locations
Storage Tier:	No storage tier has been set

Buttons: Edit, Exclude, Delete

Figure 9 The Network Map page

**Note.** When Discovery Center is run for the first time, the Network Map is empty. See page 30.



Servers identified by the application are listed in the *Discovered Locations* box by name and by total file size. To view more information about a server or location, select it in the list. Detailed information is then displayed in the *Location Summary* box.

Icons are added to the name of a location when it has specific configured attributes such as credentials, storage tier or an index. If a discovered location has been deleted or is temporarily unavailable, its folder icon is marked by a red exclamation mark.

To restore a "lost" location to the network map (and on the *Indexes*, *Areas of Interest* and *Reporting and Actions* tabs):

Click on the Reset location status button in the Location Summary.

Click on the **Show removed and inaccessible locations** check box to rebuild the network map with any "lost" locations displayed.

The page displays the following links:

- **Add Location**  
Add a location to the Network Map (see **Populating the Network Map**, below).
- **Action Selected**  
Offers a menu of actions that can be applied to all selected locations. Locations can be selected using the check boxes at the start of each row.
- **Apply tier**  
Apply or remove the tier used to calculate storage costs in the selected locations.
- **Apply credentials**  
Apply or remove the credentials to be used for the selected locations (the scheduler service credentials are used by default).
- **Delete**  
Remove the selected locations from the Network Map.
- **Exclude**  
Exclude a listed server or folder from the Network Map. If you exclude a location that has already been indexed, all information about the location will be deleted, including any indexes created in that location or its subfolders. When a location is excluded it cannot be used to configure an index.

**Note.** An AN Administrator may also exclude locations from analysis by including the relevant file paths in the Ignore Locations list when setting up the index (see Adding an index).

**Note.** Actions can also be carried out on a selected network location by clicking on the appropriate button in the Location Summary box. If you select multiple locations, summary information is not available and the summary is replaced by the Multiple Selection Summary box. This shows links to the actions available for the selected locations.

- **Collapse All**  
Close the network tree.



## Populating the Network Map

To add a server or network location to the Network Map:

1. Click on the **Add Location** link.
2. Enter the **Path**. If the address is invalid, it is highlighted in red.
3. Choose the appropriate connector from the **Location Type** dropdown list, for example: *File System*.
4. Select the **Credentials** required to access this location (see page 49).
5. Click on the **Save** button to append the location to the Discovered Locations list.

You can also use this procedure to add newly configured network locations to the network map.

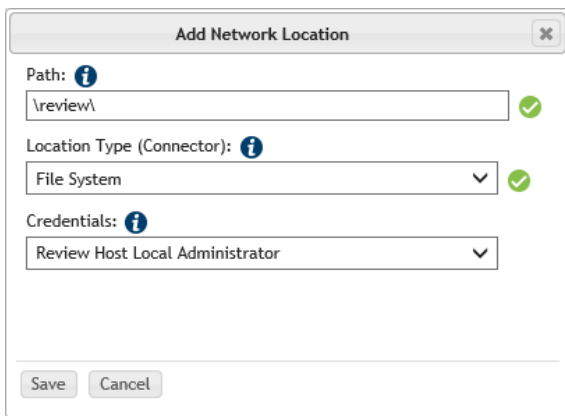


Figure 10 Adding a network location

**Note.** Locations will not be validated to confirm their presence, allowing locations to be added without needing to grant elevated credentials to the Discovery Center.

To carry out a high-level discovery of the local network environment, create an index for the new location set to *Skim* only. Add further indexes configured for deeper analysis to any interesting shares or sub-folders that are discovered during the initial skim.

You may need to restrict analysis and reporting to the required network locations by manually excluding some servers or folders.

### Removing Locations

If you want to exclude a listed server or folder from the Network Map, ensure that its associated check box is cleared (see Limiting analysis). If you deselect a location that has already been indexed, all information about the location will be deleted, including any indexes created in that location or its subfolders.

An AN Administrator may also exclude locations from analysis by including the relevant file paths in the Ignore Locations list when setting up the index (see page 77).

To remove a selected location from the Network Map, click on the **Delete Locations** button in the *Locations Summary*.

## Preventing Locations from being Indexed

By default, Discovery Center selects all discovered locations, including sub-folders, and makes them available for analysis and indexing. You may want to prevent a specific container (folder, site or document library) from being indexed:

- For convenience, to exclude system or backup folders from analysis, for example
- To respect security policy in situations where certain locations should not be made available for analysis and discovery within Discovery Center
- To maintain a tight focus on specific areas of the network

If you want to exclude one or more locations from Discovery Center discovery and indexing:

- Select the location(s) and then click on the **Exclude** button in the *Location Summary*, or
- Select the associated check boxes in *Discovered Locations* and then click on the **Action selected > Exclude** link.

Servers with one or more excluded locations are shown with a solid green check box.

**Note.** If you deselect a location that has already been indexed, all information about the location will be removed from the Discovery Center database, including any indexes created in that location or its subfolders.

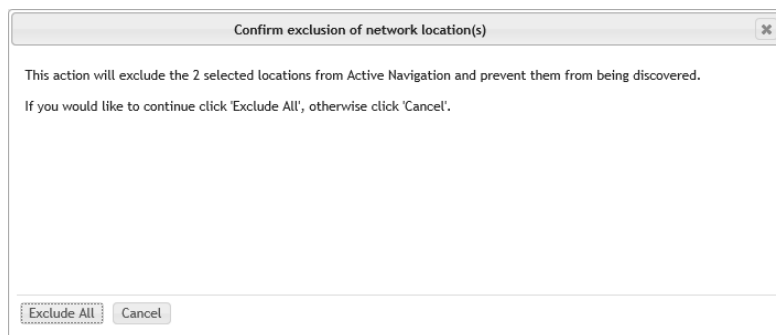


Figure 11 Excluding a location

**Note.** An AN Administrator may also exclude locations from analysis by including the relevant file paths in the *Ignored Locations* list when setting up the index (see page 94). These locations are not excluded from the network map and could be analyzed or even selected as a start location in an overlapping index (see page 78). If you create a child index, ensure that it inherits all appropriate *Ignored locations*.

## Network Map Management

The Network Map provides an overview of the topography of all repositories and locations Discovery Center has discovered; it also provides the starting point (start location) for any index; you cannot run an index without first providing a start location from the Network Map. It also provides destination locations for any actions – you cannot move or migrate files to a location unless it is included in the Network Map.

Once start locations have been configured, running an index will build the Network Map, adding locations and summary information as the index skim proceeds. Importantly, this means that if the index skim fails to find a location for any reason, it will hide that location from view because it must assume that the location is unavailable. To show locations that had previously appeared on the Network Map and were subsequently hidden, check the **Show removed and inaccessible locations** option at the bottom of the map.

## Considerations for Credential Configuration

The ability to specify credentials on the Network Map presents the opportunity to flexibly apply multiple sets of credentials to provide privileges for indexing and actioning content. However, some repositories may impose restrictions that are not fully compatible with the flexibility offered by the Discovery Center.

In particular, the mechanisms offered by Windows operating systems for access to file systems limit the use of multiple sets of credentials for concurrent access to a single file server and this can lead to unexpected errors in Discovery Center processes. This may mean that actions fail with authorization errors, or that connections to a server fail with errors of this form:

**Error connecting to remote share (error from WNetAddConnection2 : 1219 - Multiple connections to a server or shared resource by the same user, using more than one user name, are not allowed. Disconnect all previous connections to the server or shared resource and try again.)**

In order to avoid such errors, our recommendation is that you should aim to keep your configuration of credentials as simple as possible. Typically this would mean creating a specific account for Discovery Center activities and providing it with the necessary credentials to carry out your required analysis and actions.

We have constructed the Skim and Analysis processes to account for these limitations and therefore it is possible for information discovery and analysis to accommodate mixed credentials.

It has not been possible to ensure the correct behavior of action processing using mixed credentials on a single server because the processing of an action requires concurrent access to source and destination locations.

These limitations apply to repositories accessed via the *FileSystemConnector* only.





## Restoring Locations to the Network Map

If a location's status has caused it to be hidden from the Network Map, that location can be restored so that it can be used in reporting, for actions or as an index start location. To restore a location, click on it in the Network Map and then choose the **Restore to network map** option.

## Understanding when Locations are Hidden

Whilst locations may be added manually to the Network Map, its contents are driven primarily by the results of an index skim. For this reason, a Network Map location will be hidden when a skim fails to find it or does not have the credentials to access it. Note that the ability of Discovery Center to determine whether or not a location exists depends entirely upon the information returned by the relevant repository (File Share, SharePoint etc.); sometimes that information can be ambiguous due to the limitations of the programming interfaces provided by the repository vendor. Hidden locations will be marked to indicate their general status; selecting the location on the *Network Map* tab will provide further status details which are summarized below:

Table 3 Network map status codes and descriptions

Status Code	Status	Description
1	Successfully Retrieved	The resource has been skimmed by Discovery Center.
16	Discovered	The resource has been discovered but not skimmed yet.
17	Added By Admin	This resource was added manually via the Admin interface but has not been skimmed yet.
32	Path Length Exceeds Windows Limit	The path for this resource is over 260 characters long which is likely to make the file inaccessible to some applications.
33	Content Encoded Path Too Long	Some or all of the contents of this container have encoded path lengths that are too long to be recorded by Discovery Center.
34	Excluded from Analysis	The resource was excluded from analysis by Discovery Center.
48	Unauthorized	Discovery Center is not allowed to access the file or folder.
49	Not Accessible	The file or folder exists but Discovery Center is not able to read the properties.
50	Not Found	The file or folder cannot be confirmed to exist at this location
64	AN Moved Away	The resource was migrated to a new location by Discovery Center.
65	AN Archived	The resource was moved to the archive by Discovery Center.
66	AN Deleted	The resource was deleted by Discovery Center.
67	AN Ignored	The resource was ignored by Discovery Center.
68	AN Excluded	The system administrator has hidden this folder and all of its content from Discovery Center analysis.
69	AN Pending Removal	The location is scheduled to be removed from the network map.



128	Gone Away	The resource has been moved or removed by other software since it was discovered.
192	Deleted Audit Location	The resource has been deleted from the Discovery Center UI but is required by the Activity History.

## Storage Tiers

Use this tab to assign storage costs to servers in the Network Map (or shares for a file system). This allows you to track the cost of storage. Calculated storage costs are shown:

- In summary on this page in the Storage Tier table.
- On the Network Map tab in Details View panel.
- As a chart on the Reporting Overview tab.

<input type="checkbox"/>	Name ↕	Cost per GB	Usages	Total Size	Total Cost	Is Default	Actions
<input type="checkbox"/>	Tier 0 - High Performance Storage	0.00 USD	1	9.77 GB	0.00 USD		
<input type="checkbox"/>	Tier 1 - Primary Storage	0.00 USD	0	21.97 MB	0.00 USD	✓	
<input type="checkbox"/>	Tier 2 - Secondary Storage	0.00 USD	1	5.91 GB	0.00 USD		
<input type="checkbox"/>	Tier 3 - Archive Storage	0.00 USD	1	57.27 MB	0.00 USD		

Figure 12 Storage Tier tab

The tab displays the following links:

- **Add Storage Tier**  
Click on the **Add Storage Tier** link. The *Add Storage Tier* dialog box is displayed.
  1. Enter a **Name** for the Storage Tier.
  2. Enter the **Cost Per Gigabyte** (use the **Set Currency** link to select the active currency).
  3. Select the **Is default storage tier** check box if you want to apply this tier to all storage locations by default.
  4. Select a **Color** for the Storage Tier: this is used in the table on this page, Network Map and relevant reporting charts.
  5. Click on **Save**.
- **Delete Selected**  
Removes all selected Storage Tiers (chosen using the check boxes at the start of each row). Storage Tiers that were in use will be removed from all locations that they were assigned to.
- **Set Currency**  
Choose the currency to be used for Storage Tier costs. To set a custom currency, choose the **New Currency** option and then enter a **Currency Code** (the symbol to be displayed) and **Name**.
- **Reset Filters**  
Remove any filters applied to the Name column and restore the full list of Storage Tiers.



The *Storage Tiers* tab lists existing tiers under the following headings:

- **Name**  
ActiveNav includes four storage tiers by default:  
Tier 0 – High Performance Storage  
Tier 1 – Primary Storage (default)  
Tier 2 – Secondary Storage  
Tier 3 – Archive Storage
- **Cost per GB**  
Unit cost of data storage assigned to this tier.
- **Usages**  
Number of network locations explicitly assigned to this storage tier.
- **Total Size**  
Total size of network locations assigned to this storage tier.
- **Total Cost**  
Overall cost of storage for all network locations assigned to this storage tier
- **Is Default**  
Indicates if a Storage Tier is applied by default. Only one default storage tier can be defined.
- **Actions**  
Edit the properties of the selected Storage Tier (see *Add Storage Tier* above).



# System Settings

The *System Settings* page has five tabs:

- **Licensing**  
View licensing information; apply a new licensing file.
- **Users and Roles**  
Assign registered Windows users and user groups to the various Discovery Center roles.
- **Credential Management**  
Add or delete credentials for network access.
- **Discovery Center**  
Schedule index analyses and set constraints on network use.
- **Email Configuration**  
Set up an email server for the delivery of notifications.

### License Summary

Licensed To:	Active Nav								
Machine Key:	CA0F5DAF								
Expiry Date:	01 December 2022								
Solution:	Content Governance								
Feature Packs:	<table><tr><td>✓ Discovery Center</td><td>✓ Delete And Quarantine</td><td>✓ Tag And Migrate</td><td>✓ Text Analysis</td></tr><tr><td>✓ Custom Rules</td><td>✓ Automation</td><td>✓ Management Reporting</td><td></td></tr></table>	✓ Discovery Center	✓ Delete And Quarantine	✓ Tag And Migrate	✓ Text Analysis	✓ Custom Rules	✓ Automation	✓ Management Reporting	
✓ Discovery Center	✓ Delete And Quarantine	✓ Tag And Migrate	✓ Text Analysis						
✓ Custom Rules	✓ Automation	✓ Management Reporting							
Connectors:	<table><tr><td>✓ File System Connector</td><td>✓ SharePoint Connector</td><td>✓ Exchange Connector</td><td>✓ Confluence Connector</td></tr><tr><td>✓ Jira Connector</td><td>✓ OpenText Connector (Beta)</td><td>OpenText Connector</td><td>Google Drive Connector</td></tr></table>	✓ File System Connector	✓ SharePoint Connector	✓ Exchange Connector	✓ Confluence Connector	✓ Jira Connector	✓ OpenText Connector (Beta)	OpenText Connector	Google Drive Connector
✓ File System Connector	✓ SharePoint Connector	✓ Exchange Connector	✓ Confluence Connector						
✓ Jira Connector	✓ OpenText Connector (Beta)	OpenText Connector	Google Drive Connector						

*Additional connector technology by [SeeUnity](#)*

### Apply License File

To apply a valid license file please use the 'Apply License File' button.

[Apply License File](#) ⓘ

### System Usage

Volume Under Management  0% (0 B of 1.00 TB limit) is under management.

Active Navigation support or your Active Navigation account manager might request that you provide aggregated data on system usage, which you can download here. The data is not sent automatically to Active Navigation and you should email it to the address that you have been given.

The usage statistics will be packaged in an XML file and downloaded via your web browser. This can require a time consuming database query and you should avoid creating the usage statistics file while Active Navigation is busy with other work. (See the Activity tab).

If you send this information to Active Navigation it can help improve future versions of Discovery Center. See the on-line help for information about the data contained in the system usage file.

[Download Statistics](#) ⓘ

Figure 13 System Settings – Licensing



# Licensing

The Licensing tab provides information about your Discovery Center installation.

## License Summary

This section describes the currently active license:

- Licensed To**  
 Name of registered company.
- Machine Key**  
 Discovery Center uses a short key related to the hostname to lock the license. You can download a utility to run on the server to get the key before installation or it can be generated by ActiveNav after installation.
- Expiry Date**  
 Date when your current license expires.
- Solution**  
 Associated Discovery Center features are licensed collectively in Feature Packs. These can be licensed individually but, in most cases, users license Discovery Center to achieve a particular goal or *solution*. A licensed solution includes all the necessary feature packs to fulfill a particular objective. Additional feature packs may be licensed separately.

Table 4 Solutions and Feature Packs

Feature Pack	Intelligent Migration	SOLUTION Content Compliance	Content Governance
Discovery Center	•	•	•
Text Analysis*	•	•	•
Delete & Quarantine	•	•	•
Tag & Migrate	•		•
Custom Rules	•		•
Review & Automation		•	•
Management Reporting			•

\*Text analysis features enable the extraction rules configured by the deployed rule packs.



- **Feature Packs**

Your license may limit access to certain functions of Discovery Center as defined by the listed Feature Packs. Each Pack collects together logically associated features such as text analysis or file deletion. Your license should include the specific feature packs required for your own *Solution*. Unlicensed features within Discovery Center are simply hidden, greyed out or the capability is blocked. An index may fail if its configuration contains features that are no longer licensed.

All licenses include the *Discovery Center* feature pack. This allows basic Network Discovery, file property discovery and reporting on files (including duplicate analysis) and containers. Additional capabilities require the licensing of further Feature Packs as described in Table 5.

- **Connectors**

The Discovery Center uses modules, known as connectors, to allow the discovery, indexing and cleansing of files held in different sources or repositories. All licenses provide the File System Connector which supports any repository accessible using a Universal Naming Convention (UNC) path. That includes Windows File Shares and, where appropriately configured, solutions such as Novell Netware or those presenting a Windows Explorer view using WebDav protocols.

Connectors for SharePoint, Exchange, Confluence, Jira, OpenText Content Server and Google Drive are also installed but require additional licensing. The SharePoint Connector supports indexing and actions in SharePoint including the mapping of metadata for the purposes of content migration. The SharePoint and Exchange connectors support both on-premises and Office365 installations.

The Jira Connector supports indexing from Jira version 7.13.1. The Confluence Connector supports indexing from Confluence version 6.12.1. For information about other connectors, contact ActiveNav.

- **Volume Under Management**

Your license for Discovery Center includes a maximum file store size limit.

The *Volume Under Management* indicator provides information about system capacity and warns you if this limit is being exceeded. *Volume Under Management* information is also displayed on the *Indexing* page which is visible to users who are in the AN Administrators' role (see page 77).

Apply License File

Click on this button to install a new license file as supplied by ActiveNav.



Table 5 License Packs

ACTIVITY/ Feature	FEATURE PACK						
	Discovery Center	Delete & Quarantine	Tag & Migrate	Text Analysis	Custom Rules	Automation	Management Reporting
<b>INDEX</b>							
File and File Properties Discovery	•			•			
Embedded Properties Collection							
Repository Property Collection	•						
Scheduled and Repeated Discovery							
Export Index							
Import Index						•	
<b>REPORT</b>							
Report File Metadata	•						
Report Containers	•						
Report Duplicates and Masters	•						
Report Metadata Fields	•*						
Report Across Indexes	•						
Export/Import File List and Metadata		•	•				
Aggregated Reporting							
BI Tool Dashboard							
Report Metrics Over Time							•
Report Notification							•
Work Packages							•



ACTIVITY/ Feature	FEATURE PACK						
	Discovery Center	Delete & Quarantine	Tag & Migrate	Text Analysis	Custom Rules	Automation	Management Reporting
<b>ACTIONS</b>						•	
Markup Report		•	•			•	
Delete Files		•					
Quarantine Files		•					
Migrate Files			•				
Map and Write Repository Metadata			•				
Apply MIP Sensitivity Labels on Migrate			•				
Apply MIP Sensitivity Labels in Place			•				
Create Shortcuts		•	•				
Update Repository Metadata in Place			•				
Replace Characters			•				
Scheduled and Repeated Actions							
<b>ANALYZE</b>						•	
File Duplicate ID	•			•			
Content Duplicate ID				•			
File Type ID				•			
Themes and Summaries				•*			
Content Text Pattern Match							
Path Text Pattern Match				•*			
Keyword Match				•			
Proximity Match				•*			





ACTIVITY/ Feature	FEATURE PACK						
	Discovery Center	Delete & Quarantine	Tag & Migrate	Text Analysis	Custom Rules	Automation	Management Reporting
Read MIP Sensitivity Labels				•*			
<b>CLASSIFY</b>							
Hierarchical Classification	•*						
Create Classifications (allows upload of new classifications)					•		
Import Classifications	•						
Migration Mapping (enables generation of mapping CSV file)			•				
Edit Classifications (allows upload to replace existing classifications)					•		
<b>RULES</b>							
Create Extraction Rules					•		
Edit Extraction Rules					•		
Import Extraction Rules	•						
Create Fields					•		
Edit Fields					•		
Import Fields	•						

\* Requires suitable Custom Rules Pack



## System Usage

### Volume Under Management

This indicator is only displayed if your license for ActiveNav software includes a maximum file store size limit. It provides information about system capacity and warns you if this limit has been exceeded. If this occurs, it will not be possible to run an index for a new location. The Volume Under Management, if license-limited, is also displayed on the Indexing page (visible to users who are in the AN Administrators' role only).

### Download Statistics

ActiveNav support or your ActiveNav account manager might request that you provide aggregated data on system usage, which you can download from the System Usage section of the Licensing tab. The data is not sent automatically to ActiveNav and you should email it to the address that you have been given.

The usage statistics will be packaged in an XML file and downloaded via your web browser. This can require a time-consuming database query and you should avoid creating the usage statistics file while Discovery Center is busy with other work (see the Activity tab).

If you send this information to ActiveNav, it can help improve future versions of Discovery Center. The information included in the exported file is described in the following table.



Table 6 System Usage Information

System Usage Field	Description
<b>Exports</b>	Information about the installed version of Discovery Center, the version of SQL Server database, and the names of the servers that Discovery Center is installed on.
<b>Index Details</b>	Information about the Indexes that you have created: the Index names, the number of files and folders in the indexed locations, the aggregate file size and file counts within the indexed locations, and the date that the Index was last processed.
<b>Index Configurations</b>	Information about the Index Configurations used within Discovery Center: the configuration names, and the number of indexes using each configuration.
<b>Index Settings</b>	Information about the actual settings used for each index: The Index name, Skim settings, Analysis settings, and the names of the Calculated Fields used.
<b>Calculated Field Results</b>	Information about the Calculated Fields used for each index: the Index Name, the Field name, the number of matching documents, and the total size of the matching documents. (Calculated Field values are not part of the system usage).
<b>File Totals</b>	Information about the number of files in each Index: the total file count and aggregate file size in each index location.
<b>Container Totals</b>	Information about the containers in each Index: the total number of containers in each indexed location, and the number of empty containers.
<b>Index Duplication</b>	Information about the number and aggregated file size of unique and duplicated documents in each indexed location
<b>Modified Dates</b>	Information about the length of time since the files in each index were last modified (banded in to ranges relative to the time of export).
<b>Creation Dates</b>	Information about the length of time since the files in each index were created (banded in to ranges relative to the time of export).
<b>Files by Extension</b>	The total count and aggregated file sizes for Filetypes by Extension for each each index. (Filetypes by Extension is a default classification in Discovery Center).
<b>File Sizes</b>	The total count and aggregated files sizes for the size ranges in the File Size classification for each Index. (File Sizes is a default classification in Discovery Center).
<b>Task History</b>	A list of all Index names and Action types that have been processed by the ActiveNav Scheduler service, including the date and time that they were started, the time they were completed, and the total duration.
<b>Task Process History</b>	A list of all sub-tasks processed by the ActiveNav Scheduler including Index names, action types, the date and time started, the date and time completed, the duration, and the number of files actioned, and the number of exceptional items (warnings) encountered during processing.
<b>Total Duplication</b>	The total duplication across all files in the database.
<b>Extension Count</b>	The top 100 most common file extensions (by file count) that have been indexed by Discovery Center, including the total count and aggregated file sizes.
<b>Extension Size</b>	The top 100 largest file extensions (by aggregate file size) that have been indexed by Discovery Center, including the total count and aggregated file sizes.
<b>Query Execution Times</b>	The length of time needed to gather the usage statistics.



## Permissions and Access

Note that whilst a connector provides the logic necessary to work with files stored in a given repository, the Discovery Center Scheduler service or relevant Network Map location must have the privileges necessary to access or action that content. The following table shows the permissions required.

Table 7 Permission required for various tasks

Capability Required	Credentials Used	Repository Permissions Required	
		File Share	SharePoint
Basic Index Access (skim and/or analysis, required for all other capabilities)	Network Map location or scheduler	File read <sup>1</sup>	Browse directories, Use Remote Interfaces <sup>2,3</sup>
Delete files or move files from	Network Map location or scheduler	Modify	Delete items
Move files to	Network Map location or scheduler	Write	Add items
Create folder	Network Map location or scheduler	Modify	Add items
Create document library	Network Map location or scheduler	n/a	Add items
Create SharePoint column			Use Remote Interfaces
Read metadata definitions	Network Map location or Discovery Center Web interface	n/a	Use Remote Interfaces

1. Windows requires that read access be granted to the index start location and all of its parent folders for content to be successfully retrieved
2. In order to skim SharePoint access is required for all sites/site collections in the URL specified for an index. For example a SharePoint URL such as <http://server/sites/sitecollection/site> may be made up of the following levels which each require at least Basic Index Access level permissions for successful operation of the SharePoint Connector:
  - <http://server/>
  - <http://server/sites/sitecollection>
  - <http://server/sites/sitecollection/site>
  - **Default site collection**
  - **Site collection**
  - **Site**
3. SharePoint servers which do not have a default site present at root level (i.e. at <http://server/>) are not currently supported.



## File system vs Share permissions

Windows file servers allow permissions to be specified on the file system itself, and on shares set up to provide network access to content.

The permissions from Table 7 are the minimum requirement and must be configured at both share level and the file system. An overly restrictive permission set on a file share may lead to *Access Denied* errors even though the file system permissions are correctly set for the relevant user account.

A recommended approach is to set unrestrictive permissions at the share level and control access via file system permissions only. This can be an easier configuration to validate and maintain.

## SharePoint Content Selection

When a SharePoint location is indexed, the SharePoint connector focuses on lists where document content is expected. Currently the targeted list types are these document library types:

- Document Library
- Personal Document Library
- Private Document Library
- Record Library

In addition, libraries that are expected to contain content that relates to the structure of the SharePoint site are filtered out by name. Library names that are ignored are:

- Reporting Templates
- Site Assets
- Site Collection Images
- Site Pages
- Style Library

## Users and Roles

Allocating a Windows user or user group to a Discovery Center role provides that user with access to the Discovery Center interface and capabilities appropriate to their role.

Table 8 shows the functionality available to each role: note that any user may be in more than one role. The User Access page allows System Administrators to:

- Define user and group roles to limit the permitted activities of users according to the various Discovery Center roles: AN Administrator, System Administrator, Information Manager and Reviewer.
- Create user accounts and associate users with specific roles.



## Role Mapping

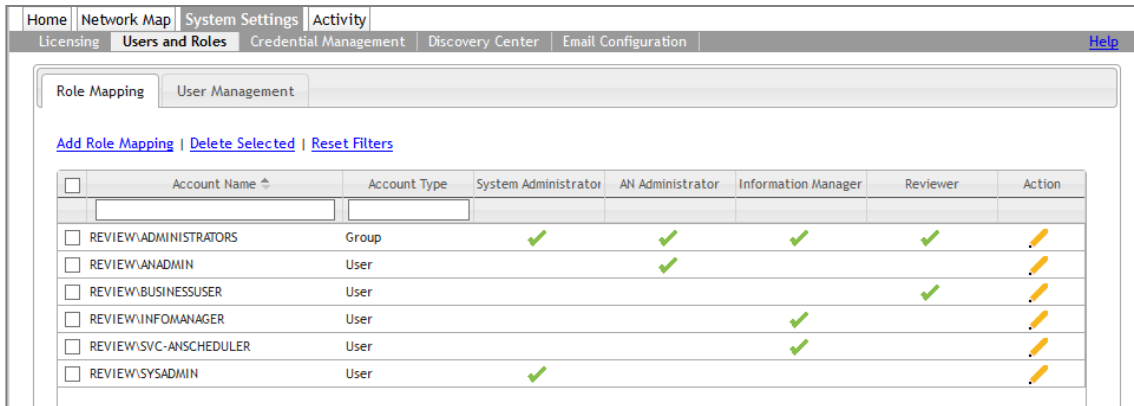
Use the Role Mapping tab to assign user roles from one or more of the available Discovery Center roles: AN Administrator, System Administrator, Information Manager and Reviewer to a Windows user or group. Use the links at the top of the Role Mapping tab to:

- **Add Role Mapping**
  1. On the *Add/Edit Role Mapping* dialog box, select the type of Windows account you want to assign to the role: **User** or **Group**.
  2. Type the Windows user account or group name into the **User/Group Name** box.
  3. Choose the Discovery Center roles to be assigned to the user or group: *AN Administrator, System Administrator, Information Manager* and *Reviewer*.
  4. Click on **OK**.
- **Delete Selected**

Delete all selected role mappings (selected using the check boxes at the start of each row).
- **Reset Filters**

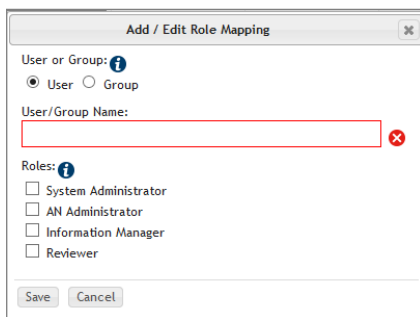
Remove any filters applied to the Account Name and/or Account Type columns and restore the full list of roles.

Click on the 'pencil' icon at the end of each row to edit the settings of an existing role mapping.



<input type="checkbox"/>	Account Name	Account Type	System Administrator	AN Administrator	Information Manager	Reviewer	Action
<input type="checkbox"/>	REVIEW\ADMINISTRATORS	Group	✓	✓	✓	✓	
<input type="checkbox"/>	REVIEW\ANADMIN	User		✓			
<input type="checkbox"/>	REVIEW\BUSINESSUSER	User				✓	
<input type="checkbox"/>	REVIEW\INFOMANAGER	User			✓		
<input type="checkbox"/>	REVIEW\SVC-ANSCHEDULER	User			✓		
<input type="checkbox"/>	REVIEW\SYSADMIN	User	✓				

Figure 14 System Settings – Users and Roles – Role Mapping tab



**Add / Edit Role Mapping**

User or Group: **?**  
 User  Group

User/Group Name:

Roles: **?**  
 System Administrator  
 AN Administrator  
 Information Manager  
 Reviewer

Save Cancel

Figure 15 Add role mapping

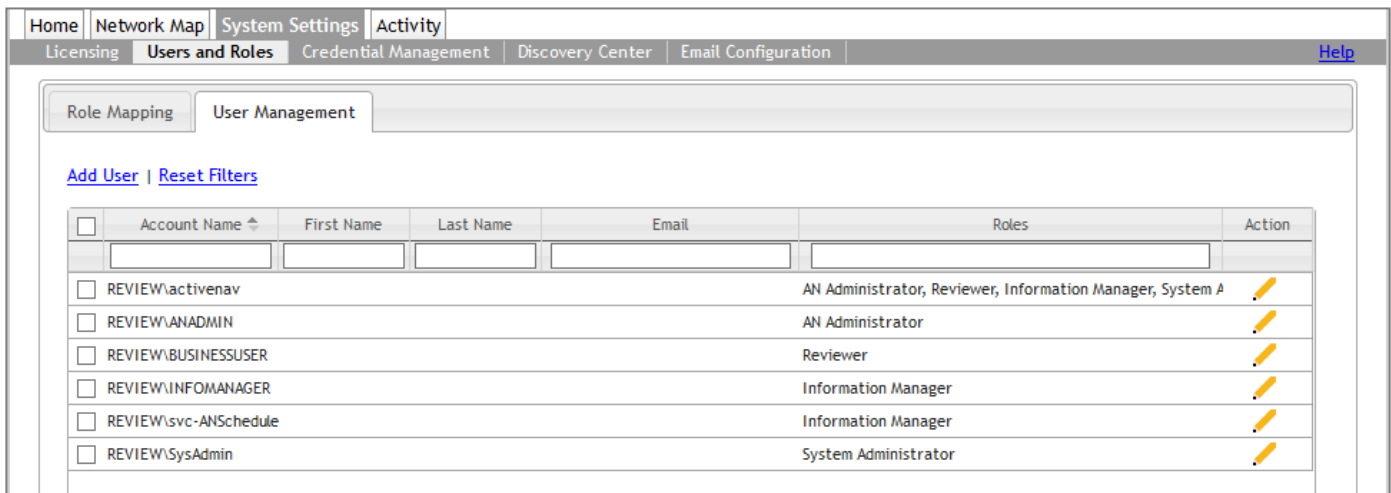
## User Management

This tab lists current system users showing name, email and permitted roles.

Use the links at the top of the tab to:

- **Add User**  
Type the new user's account name, first and last names and contact email address.
- **Reset Filters**  
Remove any filters applied to the columns and restore the full list of users.

Click on the 'pencil' icon at the end of each row to edit the settings of an existing role mapping.



<input type="checkbox"/>	Account Name	First Name	Last Name	Email	Roles	Action
<input type="checkbox"/>	REVIEW\activenav				AN Administrator, Reviewer, Information Manager, System A	
<input type="checkbox"/>	REVIEW\ANADMIN				AN Administrator	
<input type="checkbox"/>	REVIEW\BUSINESSUSER				Reviewer	
<input type="checkbox"/>	REVIEW\INFOMANAGER				Information Manager	
<input type="checkbox"/>	REVIEW\svc-ANSchedule				Information Manager	
<input type="checkbox"/>	REVIEW\SysAdmin				System Administrator	

Figure 16 System Settings – Users and Roles – User Management tab

Table 8 System Settings – User Access

Tab and Functions	Available Functions	Discovery Center Role			
		System Administrator	AN Administrator	Information Manager	Reviewer
<b>Home</b>	Provides information only according to role	•	•	•	•
<b>Network Map</b>	Manage containers for indexing or action	•			
Storage Tiers	Manage storage tiers with assigned currencies	•			
<b>System Settings</b>	Globally configure Discovery Center	•			
Licensing	Licensing and system usage information	•			
Users and Roles	Map Windows users to Discovery Center roles	•			
Credential Management	Manage Network Credentials	•			
Discovery Center	General application settings	•			
Email Configuration	Manage Email server settings	•			
<b>Metadata</b>	Discovery Center rule management		•		
Calculated Fields	Define calculated fields and map available extraction rules		•		
Extraction Rules	Define extraction rules for use in calculated fields		•		
Classifications	Import and export Classifications		•		
Metadata Import	Import from a CSV to update calculated field values		•		
<b>Indexes</b>	Define indexes for skim and analysis		•		
Indexing Overview	Manage and schedule index tasks		•		
Index Configuration	Manage index configuration options		•		
<b>Activity</b>	Review current and past activities	•	•		
Current Activity	Understand and manage tasks that are already in progress or queued	•	•		
Activity History	Review and troubleshoot completed tasks	•	•		
<b>Reporting and Actions</b>	Configure and run reports		•	•	•
Reporting Overview	Overview statistics for analysis results			•	•
Saved Views	Explore, manage and use saved views to create reports			•	•
Actions	Details of previously applied actions			•	•
Work Packages	Define or review Work Packages			•	•





Report Viewer	View existing or custom reports	•	•
Custom Queries	Execute custom database queries	•	
Mapping Rules	Define relationships between calculated fields and SharePoint metadata columns; substitutions for illegal characters.	•	
Reporting Settings	Manage the way the reporting database is kept up to date	•	

## Credential Management

Discovery Center uses credentials assigned to network map locations in order to discover, analyze and action files. If these credentials are not provided, Discovery Center uses the credentials provided to the ActiveNav Scheduler service as detailed in the Installation Guide.

Alias	Description	Usages	Actions
sm1	Credentials for sm	0	
SM2	Invalid! - No certificate file has been uploaded	0	

Figure 17 System Settings – Credential Management

Use the links at the top of the *Credential Management* tab to add or delete credentials or reset any filters applied to the list of defined credentials. Icons at the end of each row allow you to carry out actions on an existing credential:

Edit credential

View Indexes accessible using this credential

Add

1. Select the *Credentials Type*: **Username and Password**, **Certificate File**, **CyberArk AIM** or **Azure App Certificate**.
2. Type the name to be used to identify this credential in the *Alias* box.
3. Optionally, add information about the credential in the *Description* text box.
4. Enter additional information according to the selected *Credentials Type*:
  - **Username and Password**  
Enter the *Username* and *Password*.
  - **Certificate File**  
Enter the *Certificate Password*.  
Discovery Center supports PFX certificate files, which contain a certificate and private key, protected by a password.  
<https://technet.microsoft.com/en-us/library/cc770735.aspx>  
[https://en.wikipedia.org/wiki/PKCS\\_12](https://en.wikipedia.org/wiki/PKCS_12)
  - **CyberArk AIM**  
Enter the following information:



- **Safe**  
Identify the *Safe* within the CyberArk vault where credentials are located (required field).
- **Folder**  
Optional field to identify a folder within a CyberArk Safe. This can incorporate many levels with the use of a '/' character, for example: *folder1/folder2/folder3*. If this field is not populated the CyberArk Root folder for the given Safe is used by default.
- **Object**  
Required field that identifies the individual object within the given CyberArk folder.
- **Azure App Certificate (See Install Guide for details on Azure App registration requirements for this)**  
Enter the following information:
  - O365 Client/Application ID  
This is the unique value used to identify the Azure AD Application used for authentication to SharePoint Online.
  - O365 Tenant ID  
This is the unique value used to identify the Azure AD Tenant for authentication.
  - Certificate Password  
Discovery Center supports PFX certificate files, which contain a certificate and private key, protected by a password.  
<https://technet.microsoft.com/en-us/library/cc770735.aspx>  
[https://en.wikipedia.org/wiki/PKCS\\_12](https://en.wikipedia.org/wiki/PKCS_12)

5. Click on **Save**.

If you have selected a *Certificate File* or *Azure App Certificate* credential type, you will be prompted to upload it.

Delete Selected

Delete all selected credentials (selected using the check boxes at the start of each row).

Reset Filters

Remove any filters applied to the Alias and/or Description columns and restore the full list of credentials.



# Discovery Center

The *Discovery Center* tab is divided into three sections:

- Scheduling Constraints
- Global Settings
- MIP Settings - this section is only available when your license supports Microsoft Information Protection features (see **Licensing**)

## Discovery Center

Home | Network Map | System Settings | Metadata | Indexes | Activity | Reporting and Actions

Licensing | Users and Roles | Credential Management | **Discovery Center** | Email Configuration

### Scheduling Constraints

Indexing and other network intensive tasks cannot be carried out during the following times.

[Add Schedule Constraint](#)

No schedule constraints have been defined.

Times displayed in (UTC+00:00) Dublin, Edinburgh, Lisbon, London (GMT Summer Time)

### Global Settings

Maximum number of skim threads	5	
Maximum number of threads	5	
Maximum number of values for a field	100	
Export location	C:\dev-install\Exports\	
Show Disclaimer	false	
Enable Custom Queries	false	
VUM Warning (Percentage)	90	

[Edit](#)

### MIP Settings

MIP settings have been successfully validated.

If the MIP settings are updated, the ANScheduler service must be restarted for the changes to be applied.

MIP 0365 Tenant ID	5123ad5a-ab71-78f1-82b1-abc6a89e4a0e	
MIP 0365 Tenant Locale	en-US	
MIP 0365 App ID	12ac34fe-56a7-89ab-dc08-e9f62805g224	
MIP 0365 App Name	ActiveNav MIP Labelling PoC	
MIP 0365 App Version	1.0.0	
Discovery Center Credential for MIP	MIP User	
MIP 0365 Cloud Type	Commercial	

[Edit](#)

Figure 18 System Settings – Discovery Center



## Schedule Constraints

The Discovery Center Scheduler manages the running of indexes, actions and other tasks at pre-determined times or according to their progress in the task queue. However, regardless of users' desire to have work done, there may be a need to prohibit work from being scheduled at certain times – for example, to avoid a scheduled backup or virus scan. Once added, no scheduler activity will take place during schedule constraint times. Index or action tasks running at the time will be suspended until the constraint has passed.

Defined constraints are listed under the headings: *When*, *Restricted Times* and *Reason*.

Click on the **Remove** link to delete a constraint or the **Edit** link to change the schedule.

To create a scheduling constraint:

1. Click on the Add Schedule Constraint link. The Add Scheduling Constraint dialog box is displayed.
2. Type a reason for this constraint, for example: network maintenance.
3. Use the From and To end of boxes to specify the date and time of the restriction.
4. Choose how the constraint is to be applied:
  - Does not repeat
  - Every day
  - Every working day (Mon-Fri)
  - Weekly (every Thursday)
  - Monthly (on day 14)
  - CustomEnter the frequency and repeating period: days, weeks or months.
5. Click on the **Save** button.

Reason for constraint? e.g. Working hours  
Maintenance ✓

When should constraint be applied?

From: 2015/02/28 00:00 All Day

To end of: 00:00

Restriction starts on Saturday, 28 February 2015

Is the constraint applied once only or does it repeat?  
Does not repeat

Save Cancel

Figure 19 Add Schedule Constraint dialog box

## Global Settings

There are six system settings:

### Maximum number of skim threads

The maximum number of threads that the Discovery Center can use on the local machine during skimming. The performance impact will depend upon the characteristics of the environment and will be more noticeable for slower file servers. The default value is 5.

### Maximum number of threads

For a given hardware and network configuration, the *Maximum number of threads* setting controls action and analysis performance. As a rule of thumb, where the Discovery Center is the only application running on a server host, two threads can be configured for each available processor core. However, it is important to take time to determine the optimum setting for any given instance; do this by repeating the analysis of a sample data set with different thread settings and comparing performance recorded in the analysis log file. As threads are added there will be a maximum number after which performance will deteriorate.

The default value is 5. To change the value, click on the **Edit** button.

Changing the number of threads used by the Discovery Center can either limit or improve its analysis and action performance. The impact of these changes depends entirely upon the specification of the host server and, specifically, the number of available processor cores and amount of available RAM.

This setting is only implemented at the commencement of an analysis or action task. Any changes to the value will not affect a process that is already into its analysis phase. A new value will be used if the process was running but in the skim phase at the time of the change.

To apply a new maximum thread limit to an on-going analysis:

1. On the *Current Activity* page (see page 100), stop the task.
2. Change the **Maximum number of threads** value as required.
3. In the Index Configuration (see page 86) for the stopped index:
  - Disable the *Always re-analyze* option.
  - Disable the *Always re-skim* option.
  - By doing this you ensure that the index does not repeat the skim or analyze any of the files that have already been processed.
4. Restart the process.

Upon completion of the analysis, you may want to restore the original index settings for the *Always re-analyze* and *Always re-skim* options.



## Analysis File Caching

Whilst analysis is in progress, the files to be analyzed are stored temporarily in the Discovery Center file cache. The file cache is located at `Active Navigation\Discovery Center\FileCache`. Typically, one file, and its converted HTML or text counterpart, will appear in the file cache for each available analysis thread; files are removed from the file cache once the analysis of each file has been completed.

There may be a delay between the completion of analysis and all files being cleared from the file cache. If files are not being cleared from the file cache it is highly likely that some other process is holding those files open and preventing their deletion; this issue is usually caused by applications such as Windows error reporting or virus scanners.

## Maximum number of values for a field

This setting defines the limit on the number of values that can be stored against one document for any given Calculated Field. This prevents badly constructed Extraction Rules, or documents with unusual numbers of matches, from causing large numbers of values to be stored.

## Show Disclaimer

When selected (*true*), Discovery Center displays a warning screen whenever a user logs on to the system. The text and appearance of this screen can be edited by modifying `disclaimer.html`, which is located in the root location of the web UI installation. By default the presentation of a disclaimer is not enabled.

## Enable Custom Queries

Allow AN Administrators to run the defined Custom Queries during Reporting (default: *false*).

## MIP Settings

The settings in this section are required to allow communication with a Microsoft 365 instance where MIP Sensitivity Labels have been defined and published. When Discovery Center either reads an existing MIP Sensitivity Label from a file, or applies a new one, it must integrate with O365 to retrieve and validate the values of the labels on the files in relation to the set of MIP Sensitivity Labels defined for the O365 instance that have been published and are available for the given user account.

### MIP O365 Tenant ID

This is a GUID value that uniquely identifies the O365 instance - it should be available in the Azure Active Directory settings within O365.

### MIP O365 Tenant Locale

The locale value for the O365 instance - the default value for this is *en-US*.

### MIP O365 App ID

This is a GUID value that uniquely identifies the Discovery Center application that's registered in Azure Active Directory within O365. See the Installation Guide for more information on how to perform this registration and retrieve the App ID.



#### MIP O365 App Name

The name given to the Discovery Center application that is registered in Azure Active Directory within O365. See the Installation Guide for more information on how to perform this registration and retrieve the App Name.

#### MIP O365 App Version

The version of the Discovery Center application registration in Azure Active Directory within O365. The default value for this is *1.0.0*.

#### Discovery Center Credential for MIP

The credential record from Discovery Center to use as the account that integrates with O365 when retrieving and applying MIP Sensitivity Labels. The list to choose from comprises all non-certificate Credential records that have been added through the **Credential Management** page.

#### MIP O365 Cloud Type

The Azure cloud environment type where the MIP Labeling Policy has been published (for example 'Commercial' cloud or 'Government Community Cloud'). This controls the endpoints that Discovery Center will use for interacting with Microsoft Information Protection.

Unless your Microsoft 365 instance is hosted on a Sovereign Cloud, this can be left on the default value of 'Commercial'.



# Email Configuration

To set up an email server for the delivery of notifications:

1. Click on the **Edit** button.
2. Select **Enable delivery of email notifications**.
3. Enter the **Server Type**: *Exchange* or *SMTP*.
4. Type the *Host Address*.
5. Type a valid *Email Address* with access to the specified server.
6. Enter a password for the account.
7. If configuring an SMTP server, set the port number and SSL state to match your server

Click on **Save**. Alternatively, click on **Save & Test Configuration** to send an email to a specified address to test the configuration.

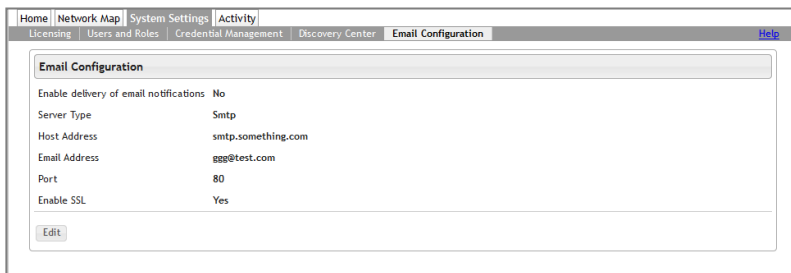


Figure 20 System Settings page – Email Configuration tab



# Metadata

**Note.** The process of defining rules to extract metadata values from your data requires detailed understanding of your requirements and the way that metadata values are found in your files. Please contact ActiveNav for further advice.

The *Metadata* page is arranged into four tabbed sections:

- Calculated Fields
- Extraction Rules
- Classifications
- Metadata Values Import

The screenshot shows the 'Metadata' page with the 'Calculated Fields' tab selected. The main table lists various calculated fields with columns for Name, Type, Classify, Report, Markup, and Fixed. A details panel on the right shows the configuration for 'A Markup Field'.

Name	Type	Classify	Report	Markup	Fixed
<input type="checkbox"/> A Markup Field	Manual markup	✓	✓	✓	
<input type="checkbox"/> Age By Created Date	All values classification		✓		✓
<input type="checkbox"/> Age By Last Accessed Date	All values classification		✓		✓
<input type="checkbox"/> Age By Modified Date	All values classification		✓		✓
<input type="checkbox"/> CLN PII	All values classification		✓		
<input type="checkbox"/> CLN Remediate	All values classification		✓		
<input type="checkbox"/> CLN ROT	All values classification		✓		
<input type="checkbox"/> CLN Stray	All values classification		✓		
<input type="checkbox"/> Copyright statement	First matching rule	✓	✓		
<input type="checkbox"/> Credit Card Number	First matching rule	✓	✓		
<input type="checkbox"/> Credit Card Terms	First matching rule	✓	✓		
<input type="checkbox"/> Document Security Terms	All matching rules	✓	✓		
<input type="checkbox"/> Document Type	First matching rule	✓	✓		
<input type="checkbox"/> Example IM Policy Release 424	All values classification		✓		
<input type="checkbox"/> File Sizes	All values classification		✓		✓
<input type="checkbox"/> Filetypes by Extension	All values classification		✓		✓
<input type="checkbox"/> Filetypes by Format	All values classification		✓		
<input type="checkbox"/> Footer	First matching rule	✓	✓		
<input type="checkbox"/> Generic IM Policy Analysis Status	All values classification		✓		
<input type="checkbox"/> Generic IM Policy Migration Readiness	All values classification		✓		
<input type="checkbox"/> Generic IM Policy Retrieval Status	All values classification		✓		
<input type="checkbox"/> Generic IM Policy Risk Sensitive Comm	All values classification		✓		

### Calculated Field Details

Name: A Markup Field

Type: Manual markup

Available for classification: ✓

Available in reports: ✓

Available for markup: ✓

Any value allowed: ✗

Allowed choices: one two

Figure 21 Metadata page – Calculated Fields tab



## Calculated Fields

This page lists the defined Calculated Fields available for selection in indexes for metadata extraction. Calculated Fields are named items of metadata derived from extraction rules and classifications and assigned to a document. There are five types of Calculated Fields: First Value Classification, All Values Classification, Best Matching Rule, Any Matching Rules and Manual Markup.

- **First Value Classification**  
In this type of Classification field, only one value for each matching document is recorded. Although the complete set of matches for each document are identified, only the value matching the highest priority node is selected and stored for the document. Nodes higher in the structure have priority over those lower down but matches within the children of a given node will have priority over matching parent nodes.
- **All Values Classification**  
In this type of Classification field, all values for each matching document are stored. A document may be assigned to multiple classification node.
- **Best Matching Rule**  
A Best Matching Rule field adopts the first value found by applying one or more extraction rules in an ordered sequence. For example, the default Document Type field uses a *File path pattern match* to extract text values from the full file path name. The Calculated Field takes the value of the first matching text identified in the file path.
- **Any Matching Rules**  
An Any Matching Rules field extracts all values found by applying one or more extraction rules. For example, a *HR Disciplinary* field could use a *Keyword match* to extract all text values matching a number of related keywords such as: *SDDP, Disciplinary, Dismissal*.
- **Proximity Matching Rules**  
A Proximity Matching Rules field extracts values found by applying two or more extraction rules. The first rule in the Selected Rules list identifies context; for example, a rule could be used to look for the text string: "Date of Birth" (and its synonyms). A second rule, defined to identify date formats, is then applied to the following or preceding text. Having identified the context using the first rule, there is an increased likelihood that any matches are indeed dates of birth and not other events. You confine the contextual limits of the secondary rule matches by specifying the number of characters to be searched Before and After the primary match of the first rule.
- **Manual Markup**  
The Manual Markup field provides a way for Information Managers and Reviewers to apply custom metadata to files. This is achieved through the use of the *Markup* reporting action (page 160 for details). For example, reviewers could use a manual markup field to identify files for review.

### Fixed Fields

Certain types of calculated field provide important reporting functionality; these fields are therefore added to all indexes and fixed so that they cannot be removed. Where necessary, their definitions can be altered by editing and uploading the supporting classification file.

- **Age By Created Date**  
Classifies files according to their age as calculated from the Created Date. The value is applied at the time that the document is classified, and so represents the age of the document at the time of classification based on the last modified date read for the file at the time the index skim was last run.
- **Age By Last Accessed Date**  
Classifies files according to their age as calculated from the Accessed Date. The value is applied at the time that the document is classified, and so represents the age of the document at the time of classification based on the last accessed date read during the preceding skim.



**Note.** On a Windows file system, the Last Accessed Date property might be updated by background tasks (such as anti-virus scans) or not updated at all depending on the server configuration. It may not therefore accurately represent the last time a document was accessed by a user for viewing, printing or other purposes.

- **Age By Modified Date**  
Classifies files according to their age as calculated from the Modified Date. The value is applied at the time that the document is classified, and so represents the age of the document at the time of classification based on the last modified date read for the file at the time the index skim was last run.
- **File Sizes**  
Classifies files in labelled size ranges.
- **Filetypes By Extension**  
Classifies files by purpose according to their file extension. The classification rules for this field include more than 4000 file extensions. Note that approximately 25% of the extensions classify into more than one file type.

## Markup Calculated Fields

Markup calculated fields are a special type of fixed field provided to support offline review of exported file level information. Instead, the values of these fields can only be changed using the **Markup** action from the *Reporting and Actions* tab, or the metadata import function. Markup fields can either allow free text entry or can be limited to a pick-list of choices.

## Calculated Field Details

The *Calculated Fields Details* box (see Figure 21) summarizes the properties of the selected Calculated Field. Click on the **Edit** link to edit the selected field.

Use the links at the top of the *Calculated Fields* tab to:

- Delete Selected
- **Export Selected**  
Export fields (in XML format) to import into another installation of Discovery Center. The export file will include any classification structures that are required to support Classification fields that are exported.
- Add New Field (see below)
- **Import Definitions**  
Import a field definition exported from another installation of Discovery Center. Note that import will overwrite fields and classification structures with the same name as those in the import file.

## Adding a New Field

To add a new Calculated Field:

1. On the *Metadata* page, select the *Calculated Fields* tab
2. Click on the **Add new field** link. The *Calculated Field Details* dialog box is displayed.
3. Enter a *Name* for the Calculated Field. Optionally, enter a *text Description*.
4. Choose the Type:
  - **First Value Classification**  
The Calculated Field will take the highest priority value according to the rules defined in the specified classification.
  - **All Values Classification**  
The Calculated Field will take all values according to the rules defined in the specified classification.
  - **Best Matching Rule**  
The Calculated Field will take value(s) from the first matching extraction rule in the list of selected rules.



- **Any Matching Rules**  
The Calculated Field will take value(s) from all the matching extraction rules in the list of selected rules.
- **Proximity Matching Rules**  
Use two or more rules with the Calculated Field taking the value(s) of matches to the selected rule only if it is found within the stated proximity of a match to the first rule in the list. No values are returned for isolated matches.
- **Manual Markup**  
No values are assigned to this field type during analysis and it cannot be selected for a particular index. The Calculated Field will only take a value if it is supplied by an Information Manager or Reviewer using the *Markup* reporting action (see page 160).

When you select an Extraction-based Calculated Field, the *Extraction Rules Settings* box is displayed (see below).

5. Choose whether the field is available for classification, reporting and/or markup.
  - **Available for classification**  
Field is available as a facet in the Classification Workbench/Designer Calculated Field tab.
  - **Available for reporting**  
Field is added to the reporting database.
  - **Available in management reports**  
Only available for classification-based fields. When selected, the field is added to the management reporting database.
  - **Available for markup**  
Field is available for Information Managers or Reviewers to update manually using the *Markup* reporting action (see page 160).

The first two options are selected by default. Deselect these options for fields that can have long textual values that need to be migrated with documents but which will not be useful to classify documents or for reporting purposes.

6. Complete the Classification Settings/Extraction Rule Settings/Markup Settings as described in the following sections.

#### Classification Settings

For a Classification-type calculated field, choose the Classification and if required the Classification Path of interest. Classification fields require a suitable classification structure designed in Classification Workbench/Designer. The required classification structure must be published to the server before configuring the Classification Field.

The value that is stored for a classification field is dependent on the path that is configured. For example when a document matches the node 'Extensions/Compressed Files/Zip File' in a classification, varying the Classification Path will have the effect shown in the table below.

Table 9 Classification Fields: values and path settings

Classification Path	Value Stored	Description
/	Extensions/Compressed Files/Zip File	Default behavior
/Extensions/Compressed/	Zip File	With a trailing slash specified, the stored value will be the child path of the specified Classification Path
/Extensions/Compressed	Compressed Files/Zip File	If a trailing slash is not specified then the last element of the specified Classification Path is included in the stored value



If the specified Classification Path is not found in the Classification then an error will be recorded and no values will be stored for the field.

## Extraction Rule Settings

Choose the Extraction Rules Settings:




1. Select one or more extraction rules. Select a rule from the *Available Rules* list and copy it to the *Selected Rules* list by clicking on the right arrow button. You may need to create your own Extraction Rule (see page 65).
2. Add further rules as required.
3. Use the up and down arrow buttons to order the extraction rules. This is particularly important if you choose a *Best Matching Rule* or *Proximity Matching Rules* field type. With *Best Matching Rule* field, the order of the rules represents your preferences (for example, the top rule is your 'best' choice). With a *Proximity Matching Rules* field, the top rule is the 'anchor'. The remaining rules are 'secondary' rules, which only return matches within the specified character proximity (see below).
4. Choose a Match Strategy:
  - **First Values:** the first matching values are stored, to the maximum defined by the Maximum Values to Store setting.
  - **Most Common Values:** the most common values are stored, to the maximum defined by the Maximum Values to Store setting.
  - **All Values:** all matching values are stored.
5. Enter a value for the **Maximum Values to Store**. This limits the number of values recorded for each analyzed file.

This setting applies to Extraction-based Calculated Fields that use the First Value and Most Common Values **Match Strategy**. Use this setting to limit the number of values stored for fields when there may be many returned values from a document. By reducing the **Maximum Values to Store** value to, say, 10, the number of values produced by the analysis can be capped.

In addition the System Setting **Maximum number of values for a field** sets an upper limit on the maximum values to store that applies to all Extraction-based Calculated Fields, including those using the All Values match strategy (see page 54 for details).

## Markup Settings

Markup Settings determine how reviewers can use the *Markup* action on the *Reporting and Actions* page (see page 160 for details):

1. Choose from the following options:
  - **Allow any value**  
Reviewers are permitted to enter any text.
  - **Restrict to fixed choices**  
Reviewers select markup comments from your defined list of allowed terms.
2. If you select *Restrict to fixed choices*, a text box is displayed for you to create, edit and delete the list of markup choices.
  - Type an allowed comment into the text box.
  - Click on . The entry is added to the list of allowed comments.
  - Repeat this procedure to add further markup text options as required.
  - To edit an existing markup choice, click on .
  - To delete an existing markup choice, click on .



## Editing a Calculated Field

To edit a Calculated Field, select it in the list on the Calculated Fields tab and then click on the **Edit** link in the *Calculated Field Details* box. The *Add Calculated Field* dialog box is displayed.

Edit the calculated field as required. Refer to the preceding sections for more information.

## Understanding Matching Strategies and Calculated Field Types

During the analysis phase, Discovery Center extracts values from every file for each extraction rule required by the Calculated Fields specified in the index configuration.

In the Calculate and Classify phase of indexing, Discovery Center examines the extracted values and assigns them to each file according to the **Type** of Extraction-based field and the choice of Extraction Rule **Match Strategy**.

When selecting values to record, the Most Common Values Match Strategy sorts values according to the number of times a given value occurs. For the First Value and All Values Match Strategy settings, values are sorted according to the rule priority and the order of matches within the document.

Table 10 Understanding Matching Strategies and Calculated Field Types

Type	Extraction Rule Match Strategy	Calculation and Classification Result
Best Matching Rule	First Values	Discovery Center assigns the first N values found by the first valid extraction rule in the Calculated Field's list, where N is the value given in Maximum Values to Store.
Best Matching Rule	Most Common Values	Discovery Center assigns the most common N values found by the first valid extraction rule in the Calculated Field's list.
Best Matching Rule	All Values	Discovery Center assigns all the values found by the first valid extraction rule in the Calculated Field's list.
Any Matching Rules	First Values	Discovery Center assigns the first N values of all valid extraction rules in the Calculated Field's list.
Any Matching Rules	Most Common Values	Discovery Center assigns the most common N values across all the valid extraction rules in the Calculated Field's list.
Any Matching Rules	All Values	Discovery Center assigns all the values of all the valid extraction rules in the Calculated Field's list.
Proximity Matching Rules	First Values	When proximity requirements are met, Discovery Center assigns the first N values of the chosen rule value in the Calculated Field's list.
Proximity Matching Rules	Most Common Values	When proximity requirements are met, Discovery Center assigns the most common N values of the chosen rule value in the Calculated Field's list.
Proximity Matching Rules	All Values	When proximity requirements are met, Discovery Center assigns all the values of the chosen rule value in the Calculated Field's list.
Manual Markup		No values are assigned to this field type during analysis, and it cannot be selected for a particular index. Values can be entered by an Information Manager during reporting using the Markup action.



**Add Calculated Field** ✕

Name:  ✔

Description:

Type:  ✔

Available for classification i
 Available for markup i

Available for reporting i
 Available for management reporting i

**Extraction Rule Settings**

Available Rules

Selected Rules

Credit Card Terms

Visa CC Rule 1

➔
⬅

⬆
⬇

Rule Name: **Credit Card Terms**      Rule Type: **Keyword match**

Data Type: **String**

Value to store:  ✔ i

Search around a match for first rule

Characters before:  ✔ i      Characters after:  ✔ i

Match Strategy:  ✔

Maximum values to store:  ✔

Figure 22 Example of Proximity Matching Rules calculated field



# Extraction Rules

The screenshot shows the 'Extraction Rules' tab in the software interface. The left pane displays a list of rules with columns for Name and Type. The right pane shows 'Extraction Rule Details' for the 'Copyright statement' rule, including its name, description, type, and pattern details.

Name	Type
<input type="checkbox"/> Copyright statement	Content pattern match
<input type="checkbox"/> Credit card	Content pattern match
<input type="checkbox"/> Credit Card AMEX	Content pattern match
<input type="checkbox"/> Credit card terms	Content pattern match
<input type="checkbox"/> Document Security Terms	Keyword match
<input type="checkbox"/> Document Type	File path pattern match
<input type="checkbox"/> Email From	Content pattern match
<input type="checkbox"/> File names that start with a TILDE character	File path pattern match
<input type="checkbox"/> Filename	File path pattern match
<input type="checkbox"/> Filename Date	File path pattern match
<input type="checkbox"/> Filename Date 2	File path pattern match
<input type="checkbox"/> Filename Version	File path pattern match
<input type="checkbox"/> Filename Version 2	File path pattern match
<input type="checkbox"/> Footer Text	Content pattern match
<input type="checkbox"/> Header Text	Content pattern match
<input type="checkbox"/> IM Support Attorney Client Privileged	Keyword match
<input type="checkbox"/> IM Support Author property	File property
<input type="checkbox"/> IM Support Bank terms	Keyword match
<input type="checkbox"/> IM Support Converted Text	Content pattern match
<input type="checkbox"/> IM Support Copyright statement	Content pattern match
<input type="checkbox"/> IM Support Credit card	Content pattern match
<input type="checkbox"/> IM Support Credit Card AMEX	Content pattern match

**Extraction Rule Details**

Name: Copyright statement

Description:

Type: Content pattern match

**Pattern Details**

Pattern: (copyright +(©|(|c|)|&copy;)+\d{4})( "[.-]" \d{4})\*

Capturing Group Number: 0

Case Sensitive: false

[Edit](#)

Figure 23 Metadata - Extraction Rules tab

This page lists the defined Extraction Rules available for use in Calculated Fields. In Discovery Center, there are several types of extraction rules:

- **File path pattern match**  
Searches the filename and path for text matching the specified pattern
- **Content pattern match**  
Searches the file contents for text matching the specified pattern.
- **Keyword match**  
Searches the file contents for the specified keywords.
- **Embedded file property**  
Takes information from embedded Microsoft Office file properties in each document, for example, *Created Date, Author, Subject*, or EXIF values from image files such as camera settings, copyright and date/time information. This rule type can also be used to extract MIP Sensitivity Label values from supported file types where MIP integration is configured (See **System Settings** for further details of these integration configuration settings).
- **Repository property**  
Extract a named item of metadata from a source repository, for example, "Created by" in Microsoft SharePoint.

Pattern match rules use Microsoft .Net regular expression syntax for matching text within documents.

The available extraction rules are listed in the left-hand table. The Extraction Rules Details table on the right shows the properties of a selection rule. Click on the Edit link to edit the selected rule.





Use the links at the top of the Metadata Extraction Rules page to:

- Add Extraction Rule
- Delete Selected
- Export Selected  
Export rules (in XML format) to import into another installation of Discovery Center.
- Import Rules  
Import rules exported from another installation of Discovery Center. Note that import will overwrite rules with the same name as those in the import file.
- Reset Filters

## Adding a New Rule

To add a new Extraction Rule:

1. On the *Metadata* page, click on the *Extraction Rules* tab
2. Click on the **Add Extraction Rule** link. The *Add Extraction Rule* dialog box is displayed.

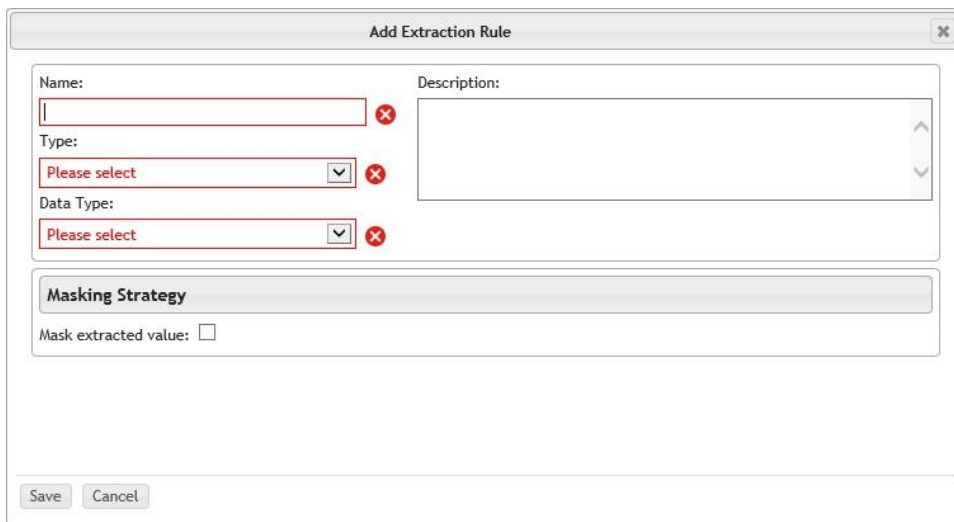


Figure 24 Add Extraction Rule dialog box

3. Enter a *Name* for the extraction rule. Optionally, enter a text *Description*.
4. Choose the *Type* (see below for details and examples):
  - **Keyword match**  
Searches the file contents for the specified keywords.
  - **File Path Pattern Match**  
Extract text from filename and path.
  - **Content Pattern Match**  
Extract text matching expression from within file content.
  - **Embedded File Property**  
Extract values from embedded Windows, MS Office, EXIF or MIP Sensitivity Label file properties.
  - **Repository Property**  
Extract a named item of metadata from a source repository, for example, "Created by" in Microsoft SharePoint.
5. Choose the **Data Type**: the format of the data to be targeted by the rule. Unless the value identified by the rule is in the correct format, nothing will be stored. The data types are:
  - String



- Integer
- Decimal
- Local Date  
Values must match the local server configuration and will be normalized to yyyy/MM/dd form. Ambiguous dates that do not match local server configuration will be rejected.
- Any Date  
Values can be any valid date format but will not be normalized as in the case of the Local Date data type.
- Credit Card  
Matched value is validated using the Luhn algorithm and separators such as white space are removed.
- IBAN Code  
Matched value must match IBAN code format and separators such as white space are removed.
- Swift Code  
Matched value must match Swift code format and separators such as white space are removed.
- National Identifier  
Choose a preformatted extraction rule pattern from the Sub Type dropdown list. For example: UK NI Number.

By choosing a specific data type, data can be extracted more efficiently and validated against the specified format.

6. Choose additional settings according to the selected **Type**:

- Keyword match (see page 67)
- File path pattern match (see page 70)
- Content pattern match (see page 70)
- Repository property (see page 73)
- Embedded file property (see page 71)

7. If required, select a masking strategy, defining how values matched by a rule can be masked before being stored in the database. This allows you to discover sensitive data, such as credit card numbers, without storing the actual value. To apply masking:

- Select Mask extracted value.
- Choose the *Part to mask*: Left, Middle or Right.
- This determines whether the extracted values are replaced by “\*” characters from the left, middle or right.
- Set the Mask size % - the percentage of the total characters in the extracted value to be replaced with “\*” characters.

The number of characters masked will be rounded up if the percentage defines a non-integer value. If the Middle option is selected and an exact partition cannot be made, the masking will extend to the right of the string.

Examples:

Original value	Mask size %	Left	Middle	Right
ABCDEF	20	**CDEF	AB**EF	ABCD**
	50	***DEF	AB***F	ABC***
12345	20	*2345	12*45	1234*
	50	***45	1***5	12***

8. Click on the **Save** button.

**Note.** If you choose a *Keyword match* with a data type other than *String*, ensure that all keywords match the selected format. If this is not the case, you will be unable to save the rule.



## Editing a Rule

To edit an existing Extraction Rule:

1. On the *Metadata* page, click on the *Extraction Rules* tab.
2. Select the rule to be edited.
3. Click on the **Edit** link in the *Extraction Rule Details* box.  
The *Add Extraction Rule* page is displayed.
4. Follow the instructions in the *Adding a New Rule* section.

## Keyword Match

Keyword extraction rules are similar to content regular expressions in that they will match an input pattern against text in the content of files, and typically will be used to look for words or phrases that can include letters, numbers and many punctuation characters.

There are two kinds of keywords: **preferred terms** and **synonyms**. Any document containing a synonym keyword is identified and reported under the preferred term, even if this does not feature in the document. Any preferred term or synonym can be flagged as case sensitive. This allows effective matching of acronyms and abbreviations. For example, with *Information Technology* as a preferred term, its acronym, *IT*, can be entered as a case-sensitive synonym.

Keywords are allowed to contain any character including spaces, for example:

**payroll num**  
**payroll #**  
**payroll:**

To be matched, the document text's 'boundaries' must be defined by two non-word characters or connector characters. For example, the keyword "payroll num" will create 'hits' for the following document text:

**\_payroll num\_**  
**payroll num**  
**.payroll num:**

but not:

**payroll number**  
**2013payroll num**

Spaces in keywords will match flexibly against text in source documents. A space character in the keyword rule will match against any number of spaces, connector characters or non-word characters (excluding tabs\*). For example, "payroll num" will create 'hits' for the following document text:

**payroll num**  
**payroll\_num**  
**payroll-num**  
**payroll.num**

\*Tab characters are intended by authors to mark a clear separation in content unlike other connector or non-word characters.

There are two ways to specify keywords: as a batch or individually.



## Adding a Batch of Keywords

If you intend to use a large number of keywords and synonyms, click on the **Add Many** link.

1. Type the keywords (preferred terms) and synonyms into the *Add Many keywords* dialog box using the following format:  
*Preferred term, 1st synonym, 2nd synonym, etc*  
using a new line for each set of keywords.
2. Precede a keyword with a caret symbol to denote that it is case-sensitive.
3. Click on the **OK** button to populate the *Keywords* section of the *Add Extraction Rule* dialog box.

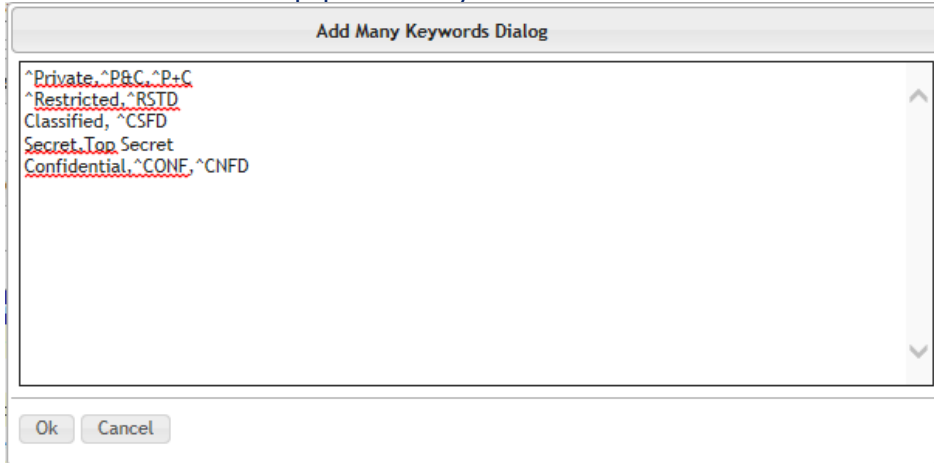


Figure 25 Adding a batch of keywords

## Adding/Editing Keywords Individually

To add a keyword:


1. Type the keyword into the top *Preferred Term* box.
2. If the keyword is case-sensitive (you want to ensure that all 'hits' exhibit the keyword precisely as you have typed it) ensure that the **Make new preferred terms case sensitive (Aa)**: check box is selected.
3. Click on the **+** button. The keyword is added to the list of preferred terms.

## Adding a Synonym

To add a synonym to an existing preferred term:

1. Select the preferred term.
2. In the *Selected Terms* box, type the keyword into the *Synonyms* box.
3. If the keyword is case-sensitive (you want to ensure that all 'hits' exhibit the keyword precisely as you have typed it) ensure that the **Make new/edited synonyms case sensitive (Aa)**: check box is selected.
4. Click on the **+** button. The keyword is added to the list of synonyms.

## Editing a Keyword

To edit a preferred term, select it in the *Preferred Terms* list, then in the *Selected Terms* box, click on the  icon alongside the preferred term.

To edit a synonym, select its preferred term, then in the *Selected Terms* box, click on the  icon alongside the synonym.

## Deleting a Preferred Term or Synonym

To delete a keyword, click on the **X** icon alongside its name in either the *Preferred Terms* or *Selected Terms* lists.

The screenshot shows the 'Edit Extraction Rule' dialog box. At the top, there are fields for 'Name' (containing 'Date of Birth'), 'Description' (empty), 'Type' (set to 'Keyword match'), and 'Data Type' (set to 'String'). Below these is the 'Keywords' section, which is divided into two panes: 'Preferred Terms' and 'Selected Term'. The 'Preferred Terms' pane has a checkbox for 'Make new preferred terms case sensitive (Aa):' which is unchecked. It contains a list of terms: 'Birth Date' (highlighted in blue), 'DoB' (highlighted in yellow), and 'Date of Birth'. Each term has a small 'Aa' icon and a red 'X' icon to its right. The 'Selected Term' pane has a 'Preferred Term' field containing 'Birth Date (Aa)' and a 'Synonyms' field which is empty and contains the text 'Preferred term currently has no synonyms'. At the bottom of the dialog, there is a 'Masking Strategy' section with a checkbox for 'Mask extracted value:' which is unchecked, and 'Save' and 'Cancel' buttons.

Figure 26 Example of a keyword-based extraction rule

The keywords you enter are actually converted to regular expressions by the analysis process. The keyword extraction rule feature simplifies the matching of basic sequences of text and, by using a keyword extraction rule, you avoid the possibility of making an error in creating a regular expression yourself.

As an example, the regular expression generated for the keyword term "payroll num" with synonym "payroll #" is:

```
(?<=(?:\p{Pc}|\W))(?:payroll(?:[\p{Pc}]|[\w\t\x0B])+num)|(payroll(?:[\p{Pc}]|[\w\t\x0B])+#))(?=(?:\p{Pc}|\W))
```

This includes the following regular expression character classes and meta characters:

- `\p{Pc}` is the set of Punctuation and Connector characters, ten characters, the most commonly used of which is the LOWLINE character (`_`).
- `\w` is the set of characters that are word characters, as specified in:  
<http://msdn.microsoft.com/en-us/library/20bw873z.aspx>.  
(Most commonly the word characters are `[a-zA-Z_0-9]`, but other Unicode characters are included in this set too).
- `\W` is the set of characters which are NOT word characters
- `\t` is a (horizontal) tab character
- `\x0B` is a vertical tab character, as often inserted by Word in bullet lists

## File Path/Content Pattern Match

If you choose File path pattern match or Content pattern match:

1. Carefully type the regular expression into the **Pattern** box.
2. Enter 0 in to the **Capturing Group Number** input to record the full matched text of the pattern, or if the regular expression contains capturing groups in parenthesis ( ) then you can specify which grouped text is to be recorded. A value of 2, for example, will return the text matching the part of the pattern enclosed by the second set of parentheses.
3. Select the **Case-sensitive** box if you want the result to match the case represented in the regular expression.

**Note.** File paths are recorded during a skim and stored in the Discovery Center database. If you make any changes to a File Path pattern rule, it is only necessary to classify the index and not to re-run it. On the other hand, if you change a content pattern rule, you will need to re-run the index (with Skim, Duplicate, and Textual settings) to analyze file content with the new extraction rule, extract any matches and classify against the Calculated Field settings.



**Add Extraction Rule**

Name:  ✓

Description:

Type:  ✓

Data Type:  ✓

---

**Pattern Settings**

Pattern:  ✓

Capturing Group Number:  ✓

Case-sensitive

---

**Masking Strategy**

Mask extracted value:

Figure 27 Example of a File Path Pattern Match extraction rule

**Edit Extraction Rule**

Name:  ✓

Description:

Type:  ✓

Data Type:  ✓

Sub Type:  ✓

---

**Pattern Settings**

Pattern:  ✓

Capturing Group Number:  ✓

Case-sensitive

---

**Masking Strategy**

Mask extracted value:

Part to mask:  ✓

Mask size (%):  ✓

Figure 28 Example of a Content Pattern Match extraction rule with a masking strategy



Edit Extraction Rule
✕

Name:  ✓

Type:  ✓

Data Type:  ✓

Description:

**File Property Settings**

Property Types:  ✓

Property:  ✓

Enter custom property name

Custom Property Name:

**Masking Strategy**

Mask extracted value:

**Figure 29** Example of a Repository Property extraction rule





## Repository Property

If you choose *Repository Property*, you can select either the "SharePoint default" or "Exchange" property set. These provide access to predefined properties for built in SharePoint columns or email message properties. In addition, for SharePoint it is possible to select "Custom" to define a non-standard property name to target a user defined column.

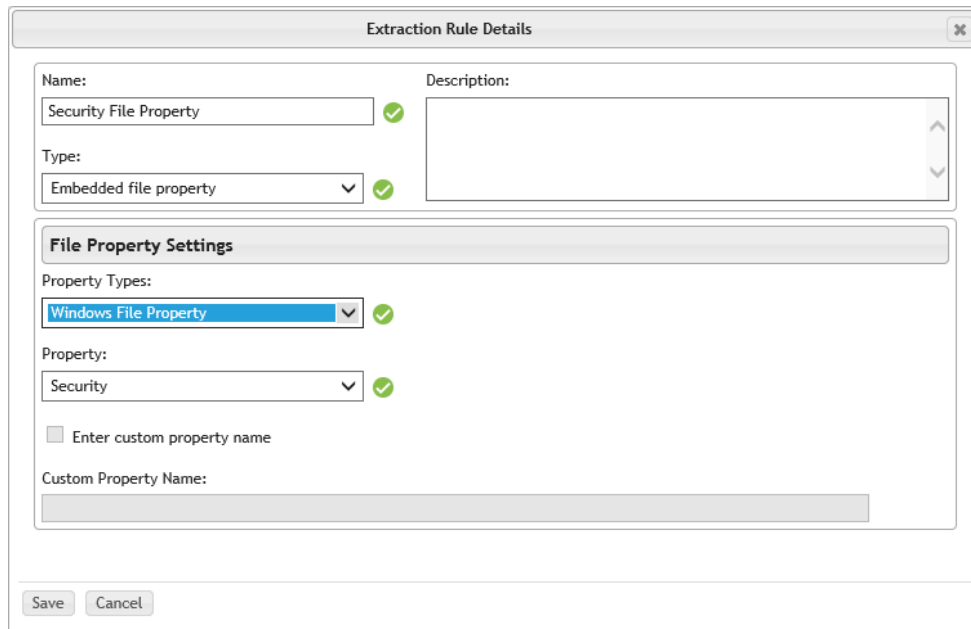
To define a Custom Property name for a connector other than SharePoint, you must select the "Other Property" Property Type. Connectors may have different conventions when specifying the Custom Property Name (e.g., case-sensitivity) that must be observed to successfully return values, due to differences in the repositories.

Refer to Appendix 8: Connector Compatibility Summary for which connectors support named properties when reading repository metadata.

## Embedded File Property

If you choose an *Embedded File property* rule:

1. Select the **Property Type** from the dropdown list:
  - Email Message Property
  - EXIF Property.
  - MIP Label Property
  - MS Office Property
  - Windows File Property
2. Choose the **Property** from the dropdown list.
3. Optionally, apply a custom name to the Property.
4. Click on the **Save** button.



The screenshot shows a dialog box titled "Extraction Rule Details" with a close button (X) in the top right corner. The dialog is divided into several sections:

- Name:** A text input field containing "Security File Property" with a green checkmark to its right.
- Description:** A large empty text area with vertical scroll bars.
- Type:** A dropdown menu showing "Embedded file property" with a green checkmark to its right.
- File Property Settings:** A section with a grey header bar containing:
  - Property Types:** A dropdown menu showing "Windows File Property" with a green checkmark to its right.
  - Property:** A dropdown menu showing "Security" with a green checkmark to its right.
  - Enter custom property name
  - Custom Property Name:** An empty text input field.
- Buttons:** "Save" and "Cancel" buttons at the bottom left.

Figure 30 Example of an Embedded Windows File Property extraction rule

## SharePoint Properties

When entering details for access to a SharePoint column you should use the internal name of the column which is not always the same as the displayed name. To check the internal name you should edit the column within the SharePoint web UI and check the name used in the Field argument of the URL.

The following built-in columns are used to populate basic metadata in the Discovery Center database and hence these are not made available for extraction as repository properties.

SharePoint Display Name	SharePoint Internal Name(s)
Created By	Author
Created	Created/Created_x0020_Date
Modified	Modified/Last_x0020_Modified

Additionally, the following built-in columns are not extracted:

SharePoint Internal Name	Notes
CheckedOutUserId	Provides internal ID, not a user name
Combine	Field with file information
FileDirRef	Available via file path regular expressions
IsCheckedoutToLocal	Can be inferred from value of CheckoutUser
_IsCurrentVersion	AN analysis always works with most recent version
RepairDocument	Link for repairing document object
Scopeld	Internal value representing list permissions
SelectFilename	Available via file path regular expressions
SelectTitle	Available via vti_title
ServerUrl	Available via file path regular expressions
_UIVersion	Version can be accessed with _UIVersionString
WorkflowVersion	Applies to workflow task objects rather than documents

The screenshot shows a dialog box titled "Extraction Rule Details". It contains the following fields and settings:

- Name:** A text box containing "Sharepoint File Type" with a green checkmark to its right.
- Description:** An empty text area.
- Type:** A dropdown menu showing "Repository property" with a green checkmark to its right.
- Repository Property Settings:** A section with a grey header containing:
  - A dropdown menu showing "SharePoint default".
  - Property:** A dropdown menu showing "File Type" with a green checkmark to its right.
  - Internal Name:** A text box containing "file\_x0020\_type".
- Buttons:** "Save" and "Cancel" buttons at the bottom left.

Figure 31 The built in SharePoint File Type property used in a repository property rule

# OpenText Content Server Properties

When entering custom property details for OpenText Content Server category attributes the Custom Property Name must be formatted as `categories[{Category Name}].{Attribute Name}`. These values are case-sensitive, so the case must match what is used in OpenText in order to return results.

For example a category named 'Legal' with an attribute called 'Notarized Date' would be formatted as: `categories[Legal].Notarized Date`.

## Classifications

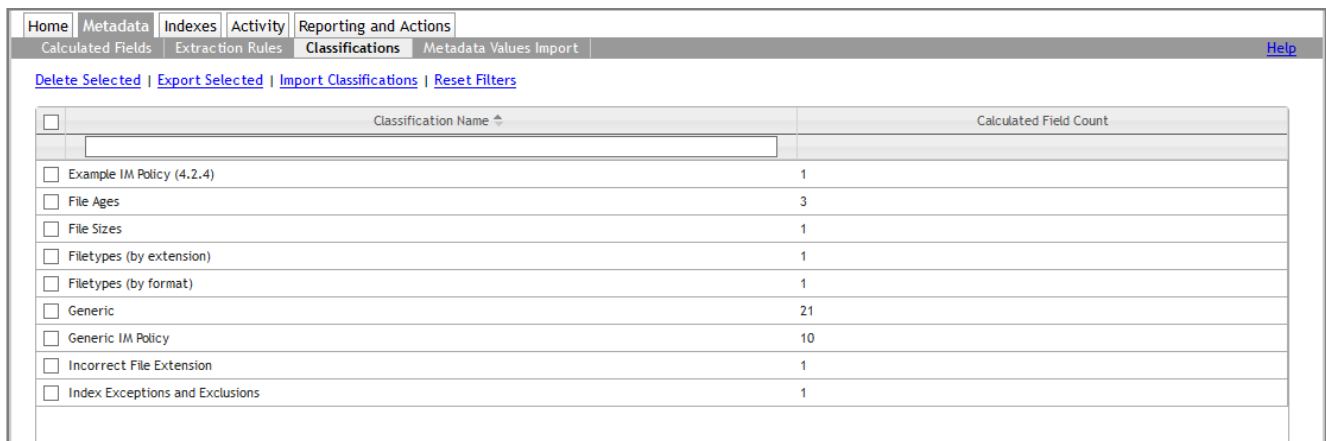
This page lists the defined Classifications by name and calculated field count (the number of calculated fields present in the classification structure). Click on the column headers to reorder the listed classifications. Click on the column header a second time to reverse the sort order.

To filter the list:

1. Type the text string into the box at the top of the *Classification Name* column.
2. Press the Return key.
3. The list is restricted to rules that contain the specified text string(s). Click on the **Reset Filter** link at the top of the page to restore the full list.

Use the links at the top of the Classifications tab to:

- **Delete Selected**
- **Export Selected**  
Export the selected classifications in XML format to import into another installation of Discovery Center or Discovery Center Workbench. By default, the exported classifications are saved in the local Downloads folder with the name: MetadataClassifications.xml.
- **Import Classifications**  
Import one or more Classifications exported from another installation of Discovery Center.
- **Reset Filter**  
Remove any filter applied to the Classification Name column and restore the full list of classifications.



<input type="checkbox"/>	Classification Name ↕	Calculated Field Count
<input type="checkbox"/>	Example IM Policy (4.2.4)	1
<input type="checkbox"/>	File Ages	3
<input type="checkbox"/>	File Sizes	1
<input type="checkbox"/>	Filetypes (by extension)	1
<input type="checkbox"/>	Filetypes (by format)	1
<input type="checkbox"/>	Generic	21
<input type="checkbox"/>	Generic IM Policy	10
<input type="checkbox"/>	Incorrect File Extension	1
<input type="checkbox"/>	Index Exceptions and Exclusions	1

Figure 32 Metadata - Classifications tab



# Metadata Values Import

To update Calculated Field values held in the Discovery Center database:

1. Exports a report in CSV format from the *Reporting and Actions* tab (see page 151).
2. Reviews and update values in the exported report using Excel or a similar tool.
3. Saves the updated report in CSV or XLSX format.
4. Import the file back into the database using the link on the *Metadata Values Import* tab of the *Metadata* page.

**Note.** For the quickest import process you should edit the import file in Excel to remove the columns for any calculated fields that have unchanged values and remove rows for documents whose values were not updated. Be careful to retain the first column (which contains the full file location) and ensure that the column headers at the top of the file match the data in the file.

## Formatting Multiple Values

When a document has multiple values for a single field then the values should be separated with two pipe characters: ||, for example

**Choice 1||Choice 2**

## Use of Quotes and Commas

Quote characters within a quoted value need to be escaped with an additional quote, for example

**”William ””Bud”” Shakespeare||Daniel ””Danny”” Boyle,Martin Scorsese”**

Exported values that contain a comma , character will be contained between quote ” characters, and this convention must also be used in files prepared for import, for example

**”red,green,blue”**

## XLSX Date and Time Formats

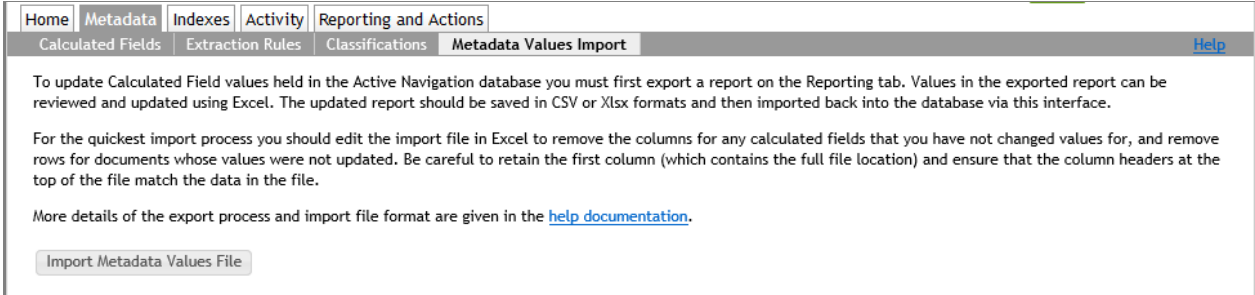
Discovery Center detects specific cell formats and processes data accordingly. However, the import library does not implement the formatting engine used by Excel thus custom formats applied in the spreadsheet will not typically be applied to the imported value.

Discovery Center applies the following rules to interpret and consistently store date, time and numeric formats:

- If a date value is found it is stored in following format: **yyyy/MM/dd**
- if a time is found along with the date then the whole format is **yyyy/MM/dd HH:mm:ss**
- If a time value is found it is stored in following **HH:mm:ss**
- If a currency or decimal or number or percent value is found then it is stored in system local format. Discovery Center inspects a currency format string to detect an appropriate currency symbol. This is the only time the Excel format template is utilized.



If it is necessary to maintain a very specific format for date/time/numeric fields when importing metadata values via an XLSX file, you must take steps to ensure that the values stored appear as text to the import process. One way of doing this is to prefix the value stored with a single quote, e.g. '2001-12-31'.



Home Metadata Indexes Activity Reporting and Actions

Calculated Fields Extraction Rules Classifications **Metadata Values Import** Help

To update Calculated Field values held in the Active Navigation database you must first export a report on the Reporting tab. Values in the exported report can be reviewed and updated using Excel. The updated report should be saved in CSV or Xlsx formats and then imported back into the database via this interface.

For the quickest import process you should edit the import file in Excel to remove the columns for any calculated fields that you have not changed values for, and remove rows for documents whose values were not updated. Be careful to retain the first column (which contains the full file location) and ensure that the column headers at the top of the file match the data in the file.

More details of the export process and import file format are given in the [help documentation](#).

Import Metadata Values File

Figure 33 Metadata - Metadata Values Import tab



# Indexes

## About Indexes

An Index is Discovery Center's basic element of content analysis, defined by a start location and a set of analysis options. Whereas a network discovery looks at the folder level, indexing extracts information from individual files, with options to identify duplication and recurrent themes and to extract metadata.

Depending on the level of analysis, document throughput can be a bandwidth and CPU-intensive procedure. It is important to schedule indexing accordingly.

The Indexes page has two tabs:

- **Indexing Overview**  
The Indexing Overview tab provides a list of predefined indexes and allows you to create, edit, schedule or delete indexes.
- **Index Configuration**  
Use the Index Configuration tab to define shared settings for indexes, comprising indexing, analysis and calculated fields for use in index creation.

## Overlapping indexes

Indexes may overlap. For example, you may want to use an index to carry out a more detailed analysis on a sub-set of the locations within a larger index. The configuration and status of overlapping indexes must be understood and carefully managed to ensure consistent results. Any 'child' index starting within a larger 'parent' index should include:

- At least the same analysis options as the parent index.  
For example, if the parent index has *duplicate analysis* configured, its child indexes must have duplicate analysis selected, as a minimum.
- The index configuration options *Always re-skim* and *Always re-analyze*.
- At least the same configured calculated fields.  
If additional calculated fields are added to the parent index, those fields should also be added to any child indexes.

Whenever you schedule the running of an index, schedule all child indexes to run afterwards to ensure overwritten results are replaced. If you delete a child index, the parent index should be run to replace deleted index results.

**Note.** It is important that all child indexes have the same security options as their parent.

## Indexing Overview

To display the Indexing Overview:

- Click on the *Indexes* tab, or
- From the *Home* page, click on the **Add Index** link.



Home Metadata Indexes Activity Reporting and Actions

Indexing Overview Index Configuration Help

Add | Update Selected | Delete Selected | Scheduling | Reset Filters

<input type="checkbox"/>	Name	Start Location	Configuration	Agent Name	Status	Files	Size	Last Complete	Scheduling	Last Modified	Actions
<input type="checkbox"/>	Test Data	\\localhost\Cloud Hosted [Default Skim, Duplicate ar -			Complete	6,541	5.91 GB	2014/12/10 11:11	Unscheduled	2014/03/14 16:16	
<input type="checkbox"/>	Personal Drive	\\localhost\Cloud Hosted [Default Skim, Duplicate ar -			Complete	1,114	1.75 GB	2014/03/21 19:19	Unscheduled	2014/03/14 16:16	
<input type="checkbox"/>	Shared Drive	\\localhost\Cloud Hosted [Default Skim, Duplicate ar -			Not Run	2,833	2.96 GB	-	Unscheduled	2014/03/14 16:16	
<input type="checkbox"/>	Live	\\localhost\Cloud Hosted [Default Skim, Duplicate ar -			Not Run	2,594	1.20 GB	-	Unscheduled	2014/03/14 16:16	
<input type="checkbox"/>	Security and ROT	\\localhost\	Security Terms	-	Not Run	6,773	5.96 GB	-	Unscheduled	2014/03/14 16:16	
<input type="checkbox"/>	G TOD test old docs	\\gjhsql2012\test old docs	Default Skim, Duplicate ar	GJHSQL2012	Imported	256	79.23 MB	-	-	2014/03/19 11:11	
<input type="checkbox"/>	Test old docs test old doc	\\localhost\test old docs	Default Skim	-	Complete	232	57.27 MB	2014/08/06 15:15	Unscheduled	2014/08/06 15:15	
<input type="checkbox"/>	TOD DupeReportTestData	\\localhost\test old docs	Generic IM Policy	-	Complete	36	11.32 MB	2014/08/06 15:15	Unscheduled	2014/08/06 15:15	
<input type="checkbox"/>	TOD UnactionedDupes	\\localhost\test old docs	Default Skim and Duplicati-	-	Complete	31	7.80 MB	2014/08/06 15:15	Unscheduled	2014/08/06 15:15	

Page 1 of 1 50

Index Action Key: Edit Index Edit Index Configuration View Imported Index Apply Index Credentials

Times displayed in (UTC) Coordinated Universal Time  
Last updated: Friday, 27 February 2015 14:46

Volume Under Management  0.6% (6.04 GB of 1.00 TB limit) is under management.

Figure 34 The Indexes page – Index Overview tab

The page displays the following commands and dropdown menus. Menu commands act on all selected indexes (selected using the check boxes at the start of each row):

- **Add**
  - Create a single index.
    - **Add Index**  
Create a single index.
    - **Add Multiple**  
Create a batch of indexes based on a single Index Configuration
- Update Selected
  - **Apply Index Configuration**  
Select an index configuration to apply to the selected indexes.
  - **Apply Credentials**  
If access to the content sources of the selected indexes requires a different network login to the one you are currently using, enter the appropriate username and password and then click on **Apply**.  
  
Select the **Remove credentials from the selected indexes** check box if you want to clear all existing network credentials from the selected indexes.
  - **Apply Security**  
Specify the users and groups permitted to access file-level results from the selected indexes. Choose from the following options:
    - **Allow all users to access file information from this index**  
Clear all existing security restrictions from the selected files.
    - Restrict the users and groups allowed to access file information from this Index  
Select the users and groups as required and then click on Save.



- **Delete Selected**

Delete all selected indexes (selected using the check boxes at the start of each row). If an index is listed in the *Current Activity* tab, either currently being processed or queued, then the index cannot be deleted.

- Scheduling

- **Process Selected Indexes**

Schedule the automatic running of the selected indexes.

**Note.** If an index is already listed in the *Current Activity* tab, either currently being processed or queued, then the index cannot be scheduled. If the index is part of a multiple selection, Discovery Center will reject the batch of indexes.

- **Recalculate Fields and Reclassify Selected Indexes**

Schedule Field Calculation and Classification for the selected indexes.

**Note.** An index cannot be scheduled for classification unless it has been skimmed. If the index is part of a multiple selection, Discovery Center will reject the batch of indexes.

- Export Selected Indexes

Queue a task to export all selected indexes (selected using the check boxes at the start of each row). Choose whether to export the full index or changes since the last export action.

For each index, an index export file with the name: ANIndexExport<IndexName>XX-xx.aie is stored at the network location specified on the System Settings > Discovery Center tab. XX represents the full version number and xx, if non-zero, the delta number for a changes only file. So, for example, ANIndexExport-GJHSQL2012-03-00.aie is the 3rd full index export file, whereas ANIndexExport-GJHSQL2012-02-04.aie is the 4th, changes only, update to the version 2 export file. Changes only exported files must be imported in the correct sequence and attempts to import an out of sequence file will be rejected. A full export can be imported at any point regardless of previous imports. A full export file contains the index, its configuration, analysis results, metadata fields (including classifications) and other database information. There is a 2GB limit on index export file size.

**Note.** It is not possible to export indexes from local network locations or from systems that contain imported indexes (export is possible if you delete all imported indexes). In addition, selected indexes must have been skimmed and not be currently active.

Export functionality is only present if the Enterprise Extensions feature is licensed.

- **Import Index**

Browse to an index export file (.aie) that has been generated on a remote server and queue a task to extract its contents. Imported indexes will be listed on the *Indexing Overview* tab with the originating servers identified by entries in the *Agent Name* column.

**Note.** You will be unable to export indexes from a system after completing an index import (unless all imported indexes are deleted). In addition, you can only import an index into the same version of Discovery Center that was used to export it and cannot import an index back into the system where it was generated. There is a 2GB limit on index export file size.

Import functionality is only present if the Enterprise Extensions feature is licensed.

- **Reset Filters**

Remove any filters applied to the *Name*, *Start Location* and/or *Index Configuration* columns and restore the full list of indexes.

The Indexes list

The *Indexing Overview* lists indexes under the following headings:





- **Name**  
Index name
- **Start Location**  
Server or folder location identified in network map for analysis by this index.
- **Configuration**  
Name of Index Configuration used to define the analysis options for this index.
- **Agent Name\***  
Name of remote server where imported index was created (blank if the index originated on this installation of Discovery Center - the central server).  
\*Note that this column is not displayed unless the index import feature has been used.
- **Status**  
View detailed status information (see below) about the index including any errors.
- **Files**  
The number of files discovered in the index start location.
- **Size**  
The amount of data discovered in the index start location (excluding content in zip files).
- **Last Completed**  
Date/time most recent analysis was completed.
- **Scheduling**  
Date/time/frequency when the index is due to be run.
- **Last Modified**  
The date that the index was last changed and saved.
- **Actions**  
Choose from the following Index management actions:
  -  **Edit Index**  
Change the options for the selected index. You can also edit the index by clicking on its name in the Indexes list.
  -  **View Imported Index** - *Imported index only*  
View imported index (it is not possible to edit an imported index).
  -  **Apply Credentials** - *Imported index only*  
If access to the content source requires a different network login to the one you are currently using, enter the appropriate username and password in the *Apply Index Credentials* box.
  -  **Edit Index Configuration**  
Switch to the *Index Configuration* page and edit the Index Configuration used by the selected index.



## Filtering

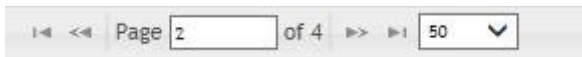
To filter the items presented in the *Indexes Overview*:

1. Type the text string into the **Name**, **Start Location** or **Configuration** boxes.
2. Press the Return key.
3. The index list is restricted to items that contain the text string. Click on the Reset Filter link to restore the full list of indexes.

## Sorting

Click on the column headers to reorder the listed indexes. Click on the column header a second time to reverse the sort order.

## Index Page Sizes



The *Indexing Overview* tab presents the indexes in pages according to the controls in the table footer. By default, each page lists up to 50 items although you can change this setting to 25, 100 or 1000 using the dropdown control. Use the other controls to browse through additional pages. Alternatively, type the page number you want to display.

Pages and index lists are automatically refreshed and sorted according to the chosen column header and ascending/descending order.

## Volume Under Management

Your license for Discovery Center includes a maximum file store size limit.

The Volume Under Management indicator provides information about system capacity and warns you if this limit is being exceeded. Volume Under Management information is also displayed on the **System Settings > Licensing** tab which is visible to users who are in the System Administrators' role (see page 36).

## Index Status

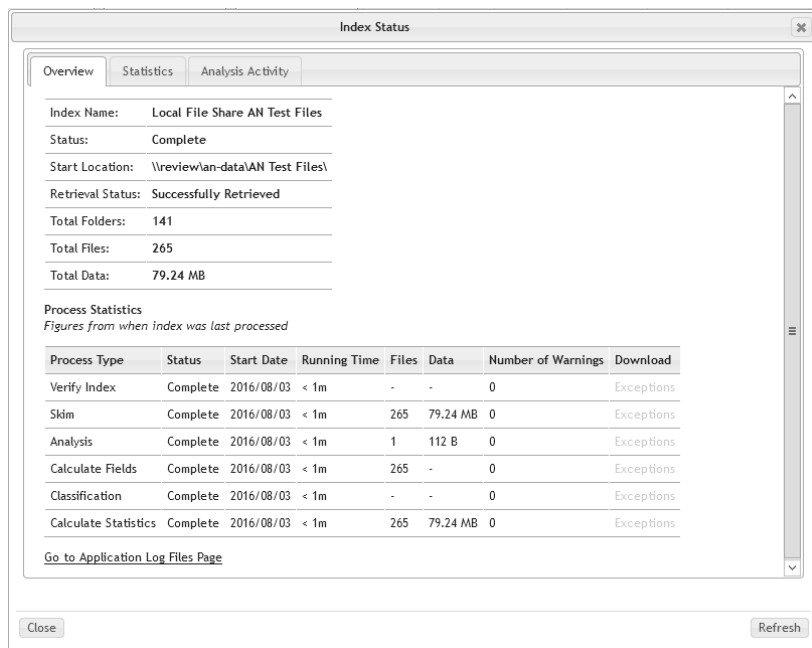
For any index listed on the *Indexes Overview*, click on its **Status** link to view detailed information about the index. The *Index Status* report is divided into three tabbed pages:

- **Overview**  
Shows information about the index and statistics from the last occasion that the index was run. This includes a *Retrieval Status* message and statistics about the index and folders, files and analysis methods. If any warnings have been issued during processing and files excluded from processing, the appropriate **Exceptions** link (in the *Download* column) is active. Click on this link to download a CSV file listing the files responsible for any warnings and more detailed error information.
- **Statistics**  
The Folder and File Statistics table shows detailed information about the number of files and folders processed during the task. The *Analysis Statistics* table lists information about the number of files analyzed for Metadata, File Format, Duplicate and Thematic content. This includes an *Analysis Status* message.
- **Analysis Activity**  
This tab provides information about an active index including the names of the files currently undergoing analysis and the



duration of the task. This can be used to aid troubleshooting if an index appears to be running slowly. Click on the **Refresh** button to update the tab.

For further troubleshooting, click on the **Go to Application Log Files** page link.



The screenshot shows a window titled "Index Status" with three tabs: "Overview", "Statistics", and "Analysis Activity". The "Overview" tab is active and displays the following information:

- Index Name: Local File Share AN Test Files
- Status: Complete
- Start Location: \\review\an-data\AN Test Files\
- Retrieval Status: Successfully Retrieved
- Total Folders: 141
- Total Files: 265
- Total Data: 79.24 MB

Below this is a section for "Process Statistics" with the subtitle "Figures from when index was last processed". It contains a table with the following data:

Process Type	Status	Start Date	Running Time	Files	Data	Number of Warnings	Download
Verify Index	Complete	2016/08/03	< 1m	-	-	0	Exceptions
Skim	Complete	2016/08/03	< 1m	265	79.24 MB	0	Exceptions
Analysis	Complete	2016/08/03	< 1m	1	112 B	0	Exceptions
Calculate Fields	Complete	2016/08/03	< 1m	265	-	0	Exceptions
Classification	Complete	2016/08/03	< 1m	-	-	0	Exceptions
Calculate Statistics	Complete	2016/08/03	< 1m	265	79.24 MB	0	Exceptions

At the bottom of the window, there is a "Close" button on the left and a "Refresh" button on the right. A link "Go to Application Log Files Page" is also present.

Figure 35 Example of an Index Status report



**Table 11** Retrieval messages

Code*	Retrieval Message	Description
1	Successfully Retrieved	The resource has been skimmed by Discovery Center.
16	Discovered	The resource has been discovered by the Discovery Center but not been skimmed yet.
17	Added By Admin	This resource was added manually via the Admin interface but has not been skimmed yet.
32	Path Length Exceeds Windows Limit	The path for this resource is over 260 characters long which is likely to make the file inaccessible to some applications.
33	Content Encoded Path Too Long	Some or all of the contents of this container have encoded path lengths that are too long to be recorded by Discovery Center.
48	Unauthorized	Discovery Center is not allowed to access the file or folder.
49	Not Accessible	The file or folder exists but Discovery Center is not able to read the properties.
50	Not Found	The file or folder cannot be confirm to exist at this location
64	AN Moved Away	The resource was migrated to a new location by Discovery Center.
65	AN Archived	The resource was moved to the archive by Discovery Center.

\* Error codes are displayed in the database schema only



Table 12 Analysis messages

Code*	Analysis Message	Description
1	Successfully Analyzed	The resource has been accessed by the Discovery Center.
2	Waiting for Analysis	The resource has not yet been analyzed but is queued for analysis.
3	Internal Server Error	The Discovery Center server has experienced an internal error that prevented analysis of the document.
4	Document Temporarily Unavailable	A temporary problem prevents access to the document for analysis.
5	Access Attempt Timed Out	The analysis was unable to read the document in the time allowed.
6	Out of Memory	The Discovery Center server tried to use more than the available memory during analysis of the document.
7	Unable to Create Temporary File	The Discovery Center server was unable to create a temporary file that is needed to analyze the document.
8	Temporary File Missing	The temporary file created by the Discovery Center server has been removed before document analysis was completed.
16	Password Protected	The document was not analyzed because it is protected by a password.
32	Corrupt File	The document to be analyzed was not in the expected format.
33	Unexpected end of file.	The document to be analyzed ended unexpectedly.
34	Unsupported Format	The document to be analyzed is not in a format that is currently supported by Discovery Center analysis.
35	No content	The document to be analyzed has no content.
48	Unknown Language	The document was not in a language supported by Discovery Center analysis. (Currently English, French and Spanish are recognized, only analysis results in English are supported.)
64	Too Big	The document is larger than the maximum size that has been specified for analysis.
65	Too Small	The document is smaller than the minimum size that has been specified for analysis.
66	Not Wanted	The document type has been excluded from analysis.
67	Regex Analysis Timeout	Regular expression processing exceeded the configured timeout.
128	Failed XPath Match	No part of the XML file matched the supplied XPath query.
129	No Category	No suitable category was found for this document.
255	Undefined error	Undefined/unhandled error.

\*Error codes are displayed in the database schema only






# Index Configuration

Use the Index Configuration tab to define index templates for use in index creation (see Adding an Index and Adding Multiple Indexes). Each configuration comprises indexing, analysis, calculated fields and security options.

The page displays the following links:

- **Add Index Configuration**  
Create a new Index Configuration
- **Delete Selected**  
Delete all selected index configurations (selected using the check boxes at the start of each row).
- **Export Selected**  
Export one or more index configurations (selected using the check boxes at the start of each row). You are prompted to choose the download location and given the option to rename the file (default name: *IndexConfigurationExport.aic*).
- **Import Configurations**  
Import one or more index configurations from an index configurations export file (extension .aic). Current index configurations will be overwritten if the .aic file contains any index configurations with the same name.
- **Reset Filters**  
Remove any filter applied to the *Name* or *Summary* columns and restore the full list of index configurations.

Previously defined index configurations are listed under the following headings. To sort by a particular property, click on the column headers. Click on the column header a second time to reverse the sort order:

- **Name**  
Index Configuration name
- **Summary**  
Brief description of Configuration including number of Calculated Fields.
- **Index Count**  
Number of indexes defined using this configuration.
- **Actions**  
Choose from the following actions:
  -  **Edit Configuration**  
Change the options for the selected Index Configuration.
  -  **Create Index from this Configuration**  
Switch to the Indexing Overview tab and add an index based on the selected Index Configuration.
  -  **View Indexes using this Configuration**  
Switch to the Indexing Overview tab, applying this Index Configuration as a filter.

## Filtering

To filter the items presented in the Index Configuration:

1. Type the text into the **Name** and/or **Summary** boxes.
2. Press the Return key.

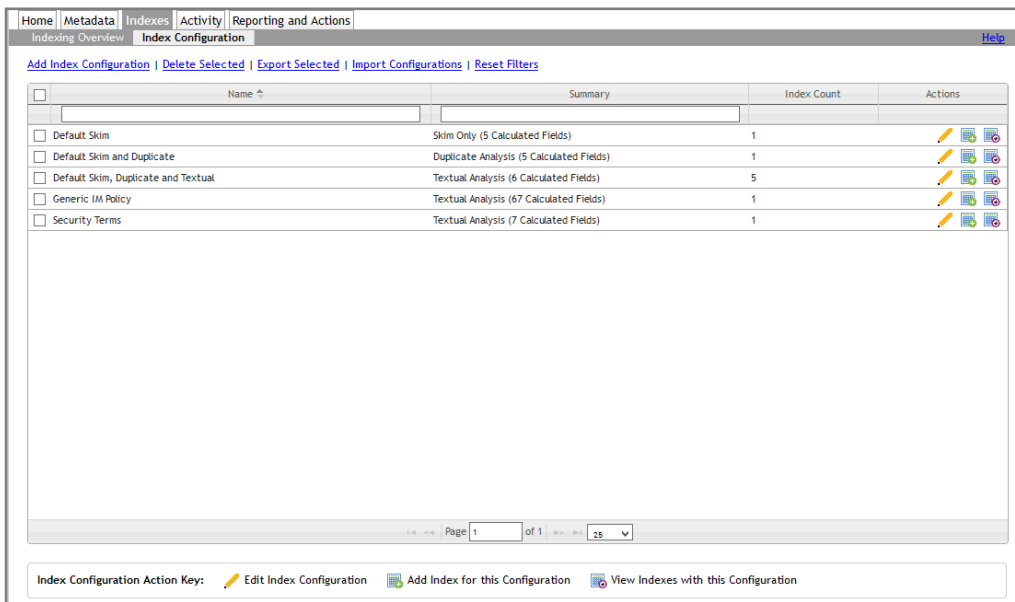
The Index Configuration list is restricted to items that contain the text. Click on the **Reset Filters** link to restore the full list of index configurations.


















## Index Configuration Page Sizes

The Index Configuration tab presents the index configurations in pages according to the controls in the table footer. By default, each page lists up to 25 index configurations although you can change this setting to 50 using the dropdown control. Use the other controls to browse through additional pages. Alternatively, type the page number you want to display.

Pages and index configuration lists are automatically refreshed and sorted according to the chosen column header and ascending/descending order.



<input type="checkbox"/>	Name	Summary	Index Count	Actions
<input type="checkbox"/>	Default Skim	Skim Only (5 Calculated Fields)	1	  
<input type="checkbox"/>	Default Skim and Duplicate	Duplicate Analysis (5 Calculated Fields)	1	  
<input type="checkbox"/>	Default Skim, Duplicate and Textual	Textual Analysis (6 Calculated Fields)	5	  
<input type="checkbox"/>	Generic IM Policy	Textual Analysis (67 Calculated Fields)	1	  
<input type="checkbox"/>	Security Terms	Textual Analysis (7 Calculated Fields)	1	  

Page 1 of 1 | 25




Index Configuration Action Key:  Edit Index Configuration  Add Index for this Configuration  View Indexes with this Configuration

Figure 36 Index Configuration tab

## Adding an Index Configuration

To add an index configuration, click on the **Add Index Configuration** link on the Index Configuration tab. To edit an existing Index Configuration, click on  in the *Actions* column. The *Add/Edit Index Configuration* page is displayed.

The *Index Configuration* page consists of three tabbed pages as described below. When you have finished, click on the **Save & Close** button to save the configuration, or the **Save & Create Index** button to save the new configuration and move directly to the *Add*

*Index* dialog box. Saved index configurations are appended to the list on the *Index Configuration* page and are available for index creation (see page 86).

Type a name to identify the index configuration in the *Name* text box. Choose a meaningful and descriptive name related to the purpose so that other users will be able to identify it easily.

Choose the Analysis Mode (the level of detail required for the index configuration - for more information, see page 97):

- **Skim**  
Information is collected about the folder structure and basic properties of files and folders, without analysis of file content.
- **Skim and Duplicate**  
This adds basic analysis of the content of each file, limited to identifying Duplicate (identical copy) documents. Duplicate analysis generates data for duplication reporting and cleansing. This is the minimum requirement for duplication analysis.
- **Skim, Duplicate and Textual** (Analysis Pack license required)  
This adds Textual (text content) analysis in addition to the previous levels of detail. Textual analysis includes File Format identification, Thematic analysis, Keyword and Text Content Pattern extraction, and File and Repository Property analysis. File Format analysis allows you to set up reports based upon the exact file format determined by inspecting file contents rather than filename extension. Thematic analysis consists of the extraction of file metadata to use as themes and/or summaries and to find similar files. It includes the keywords, description/comments and title fields from both Office style files and HTML web pages.

## Calculated Fields

The Calculated Fields tab allows you to select data fields for extraction and analysis.

To identify the calculated fields to be analyzed by indexes based on this configuration:

1. From the *Available Fields* list, select a calculated field you want to extract.
2. Any field identified by an asterisk requires *Full Content Analysis* to yield results. If you attempt to add one of these fields with *Skim Only* or *Skim and Duplicate* selected you will be prompted to *enable Skim, Duplicate and Textual*. In addition, an error message is displayed on the *Edit Index* page if one of these calculated fields is selected but the analysis type is amended to *Skim* or *Skim and Duplicate*.
3. Click on the right arrow key to add the highlighted field to the *Selected Fields* list.
4. Repeat the procedure for all required metadata fields.
5. Order the fields in the *Selected Fields* list using the up and down arrow keys.

To remove a metadata field from the *Selected Fields* list, select it and click on the left arrow button.

**Note.** The *Selected Fields* list contains standard, fixed, fields: File Sizes and File Types by Extension. These cannot be removed from the list and are always calculated during analysis.





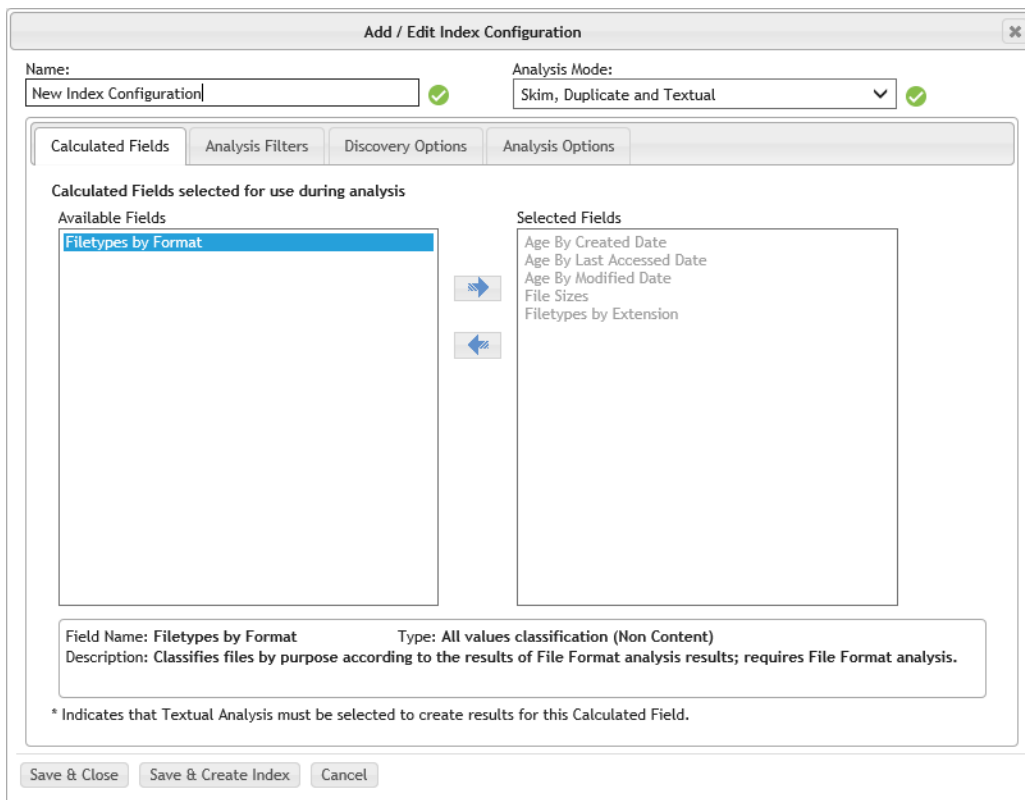


Figure 37 Index Configuration – Calculated Fields tab

## Analysis Filters

Filters allow you to improve performance by limiting analysis to the files of interest. When filters are active, analysis statistics include a count of the files excluded from analysis. The tab is grayed and disabled if you have selected the Skim analysis mode.

### Maximum File Size (MB)

Extremely large file can significantly increase the time that it takes to complete analysis for an index. Often, files of this size are log or data files, which cannot be interpreted by Discovery Center.

Use this setting to apply a maximum file size: any file exceeding this limit will be ignored by Discovery Center during the textual phase of analysis. However, the Ignore conditional filter for duplication setting (see Analysis Options, below) enables Discovery Center to ignore these restrictions for duplicate analysis. This takes place without saving files in to the server file cache and, since there is no textual analysis, it has reduced impact on the total indexing time.

By default, the file size limit is 20MB. Any files exceeded the limit are given an index status of 'Too Big' and can be highlighted in an Index Failures classification or matched by the Index Exceptions and Exclusions calculated field.

### Permitted File Extensions

You can exclude files (such as log or data files) by specifying an exclusion rule based on their file extensions. Click on the Manage extensions link to restrict analysis to specific file types. Type separate entries with a newline (return) character, for example: .doc and .docx. Analysis speed can be greatly improved if you can target analysis in this way. By default, all files are analyzed.

### Calculated Field Filter



For more advanced scenarios, you can specify a calculated field to determine which files will be analyzed: either include or exclude files that have a value for the calculated field. The field is populated after skim and determines whether a file should be analyzed. Typically, this "conditional field" will be a classification, which should be based on skim results. If it is based on a non-classification field then the field cannot be one that requires content analysis. Use of fields other than classifications may be hard to understand, for example the first analysis will not be subject to the conditional field because there are no results available. If a Calculated Field Filter is enabled, the indexing process shows an additional "Apply conditional filters" process between the skim and analysis processes.

- **Filter type**  
Choose whether to *Include* or *Exclude* files with matches to the selected field.
- **Filter field**  
Choose the calculated field to be used for the filter.

The screenshot shows a window titled "Add / Edit Index Configuration" with a close button (X) in the top right corner. Below the title bar, there are two input fields: "Name:" with the value "New Index Configuration" and a green checkmark, and "Analysis Mode:" with a dropdown menu showing "Skim, Duplicate and Textual" and a green checkmark. Below these are four tabs: "Calculated Fields", "Analysis Filters" (which is selected), "Discovery Options", and "Analysis Options". Under the "Analysis Filters" tab, there are two sections: "Basic Analysis Filters" and "Advanced Analysis Filters". In the "Basic Analysis Filters" section, "Maximum File Size (MB):" has a text input field with "500" and a green checkmark. "Permitted File Extensions:" has a text input field with the text "Files with any extension will be analyzed. [Manage extensions](#)". In the "Advanced Analysis Filters" section, there is a checked checkbox for "Enable calculated field based filter". Below this, "Filter type:" has a dropdown menu with "Include" selected and a green checkmark. "Filter field:" has a dropdown menu with "Age By Modified Date" selected and a green checkmark. At the bottom of the dialog, there are three buttons: "Save & Close", "Save & Create Index", and "Cancel".

Figure 38 Index Configuration – Analysis Filters tab



## Discovery Options

This tab provides additional options controlling discovery and analysis:

- **Always re-skim**

Always carry out a skim-level discovery of the selected location even if it has been skimmed previously. Selecting this option will ensure that files added to or removed from the indexed location are detected and accounted for during analysis.

- **Include zip file content**

Analyze files compressed within zip archives.

**Note.** The expansion of Zip files can dramatically increase the quantity of information you will record during analysis and affect the duration of the detailed analysis phase. It also requires space in the server file cache folder to expand the entire content of the archives during analysis. Before choosing this option, it is advisable to perform an initial skim to assess the number and size of Zip files on your system.

- **Ignored filename patterns**

Use this setting to indicate the patterns of any filenames to be ignored when encountered during the skim process. Patterns are specified here as regular expressions, with multiple ignore patterns being supported in a single configuration when each individual pattern is included on a separate line in the input text box. By default the ignore pattern *snapshot* is included in all index configurations.

- **Retrieve file owner**

Use this option to retrieve a file ownership property (if available) from each file in an indexed share during the skim phase. The property is collected from the file owner property in Windows file shares and the created by property from files in SharePoint. You can view the File Owner property on the Basic Metadata tab in a Report's File List view.

**Note.** Retrieving the file ownership property can reduce skim performance (depending on the state of the local environment).

- **Include admin shares**

By default, Discovery Center does not index any locations set up as Administrative Shares. Select this option if you want Discovery Center to index Admin Shares within chosen network locations.

**Note.** When Admin Shares are discovered during an index they each have an associated type, set by the Operating System, which can be any of the following:

- STYPE\_DISKTREE
- STYPE\_PRINTQ
- STYPE\_DEVICE
- STYPE\_IPC
- STYPE\_CLUSTER\_FS
- STYPE\_CLUSTER\_SOFS
- STYPE\_CLUSTER\_DFS
- STYPE\_SPECIAL
- STYPE\_TEMPORARY

With **Include admin shares** selected, only admin shares with the STYPE\_DISKTREE type will be indexed, the index logic explicitly ignores all shares of type other than STYPE\_DISKTREE to exclude operating system or similar system files in reporting.

- **Automatically generate export file**

Select this option if you want to generate an export index file (.aie) whenever an index using this configuration is processed. For a new index, a full export index file is created; for subsequent updates, an incremental, 'changes only' export file is generated. Use this feature if you need to transfer index data to a central server. Export index files are stored at the location specified on the *System Settings > Discovery Center* tab.



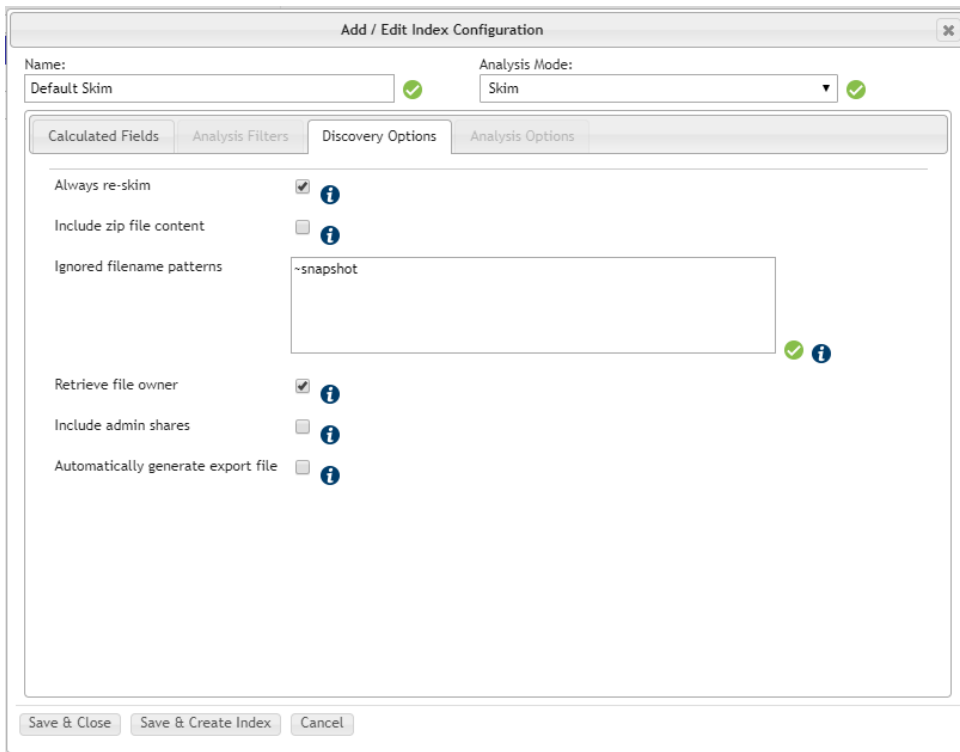


Figure 39 Index Configuration – Discovery Options tab



## Analysis Options

Analysis Options allow you to customize similar match and thematic analysis options. The tab is only displayed if you have selected the Skim, Duplicate and Textual index option.

- **Always re-analyze**  
Analyze files, even if they have been indexed by previous crawls.
- **Enable thematic analysis**  
Thematic analysis is carried out by default as part of textual analysis. It is a key part of the analysis, used to populate keywords and summaries when migrating documents in to SharePoint for example, and to extract themes used in classification rules (node search terms). If this type of analysis is not necessary, you can improve indexing performance by disabling thematic analysis.
- **Ignore max file size for duplication**  
Files exceeding the Max File Size for Analysis limit are included by default in duplication analysis. This takes place without the downloading of files across the network and, since there is no textual analysis, it has little impact on indexing performance. Clear this check box if you want to exclude large files from duplication analysis.
- **Enable content duplicate analysis**  
Content duplicate analysis is intensive in terms of processing time and is only necessary if your network contains SharePoint locations or Office files that have previously been stored in SharePoint (see Duplicates Reports). If this is not the case, you can disable content duplicate analysis to improve performance.

The screenshot shows the 'Add / Edit Index Configuration' dialog box with the 'Analysis Options' tab selected. The 'Name' field is 'New Index Configuration' and the 'Analysis Mode' is 'Skim, Duplicate and Textual'. The 'Analysis Options' section contains the following settings:

Option	Checked	Info Icon
Always re-analyze	<input type="checkbox"/>	Yes
Enable thematic analysis	<input checked="" type="checkbox"/>	Yes
Ignore conditional filter for duplication	<input checked="" type="checkbox"/>	Yes
Enable content duplicate analysis	<input checked="" type="checkbox"/>	Yes

The 'Thematic Analysis Options' section includes:

Option	Value	Valid	Info Icon
Maximum number of themes	30	Yes	Yes
Maximum percentage of themes	80	Yes	Yes
Number of summary sentences	5	Yes	Yes
Enable similar match reporting	<input checked="" type="checkbox"/>	Yes	Yes

Buttons at the bottom: 'Save & Close', 'Save & Create Index', and 'Cancel'.

Figure 40 Index Configuration – Analysis Options tab

These options are available if you have selected *Enable thematic analysis* (see above).

- **Maximum number of themes**  
The maximum number of themes to be extracted from a file during analysis (default: 20)
- **Maximum percentage of themes**  
The percentage of possible themes to include (default: 80)
- **Number of summary sentences**  
Number of sentences to include in the generated summary (default: 5)  
**Note.** This can greatly increase the run time of the index.

## Adding an Index

To create a single index, click on the **Add Index** link on the *Indexing Overview* page. The *Add Index* page is displayed.

**Note.** Click on the **Add Multiple Indexes** link if you want to create a batch of indexes based on the same Index Configuration (see page 86)

The screenshot shows the 'Add Index' dialog box. It has a title bar 'Add Index' and a close button. The dialog is divided into two main sections. The left section contains: 'Index Name' with an empty text box and a red 'x' icon; 'Index Configuration' with a dropdown menu showing 'Please select' and a red 'x' icon; 'Network Credentials' with a dropdown menu showing 'Please select'; and 'Visibility of Results' with a link 'Manage Index Security (No security restrictions)'. The right section contains: 'Start Location' with an empty text box and a red 'x' icon; a list of locations with radio buttons: 'gjhsql2012 [79.23 MB]' and 'localhost [5.96 GB]'; and a link 'Manage Ignored Locations (No locations ignored)'. At the bottom are three buttons: 'Save & Close', 'Save & Schedule', and 'Cancel'.

Figure 41 The Add Index dialog box

1. Type a name to identify the index in the **Name** text box. Choose a meaningful and descriptive name related to the starting location and purpose so that other users will be able to identify it easily.
2. Browse to the *Start Location* of the content that you want to analyze using the displayed network map. If the location is not listed contact your System Administrator.
3. To create a list of locations to exclude from indexing, click on the **Manage Ignored Locations** link. List the folders to exclude from indexing. Each location should be on a separate line. For example:  
`\\server\share\business\marketing`  
`\\server\share\business\sales`

4. Choose the Index Configuration to be used for this index. This determines the indexing, analysis, calculated fields and security options. Three default index configurations (see below) or custom configurations can be defined on the Index Configurations tab (see page 87).
    - **Default Skim**  
Information is collected about the structure and properties of files and folders without analysis of file content.
    - **Default Skim and Duplicate**  
This adds basic analysis of the content of each file, limited to identifying *Duplicate* (identical copy) documents. Duplicate analysis generates data for duplication reporting and cleansing. This is the minimum requirement for duplication analysis.
    - **Default Skim, Duplicate, and Textual** (Analysis Pack required)  
This adds Textual (text content) analysis in addition to the previous levels of detail. Textual analysis includes File Format identification, Thematic analysis, Keyword and Text Content Pattern extraction, and File and Repository Property analysis. File Format analysis allows you to set up reports based upon the exact file format determined by inspecting file contents rather than filename extension. Thematic analysis consists of the extraction of file metadata to use as themes and/or summaries and to find similar files. It includes the keywords, description/comments and title fields from both Office style files and HTML web pages (see **Metadata** on page 22 for further information about this option).
  5. Analysis normally proceeds with the credentials used by the Scheduler Service configured by the Systems Administrator (see the **Installation Guide** for details). In situations where these credentials are not sufficient to analyze one or more of the listed locations, indexing will fail unless you select appropriate credentials from the **Network Credentials** dropdown list.
  6. Click on the **Manage Index Security** link to specify the users and groups permitted to access file-level results from this index. Choose from the following options:
    - Allow all users to access file information from this Index
    - Restrict the users and groups allowed to access file information from this Index  
Select the users and groups as required.
  7. Save the Index:
    - Click on **Save and Schedule** if you want to run the index immediately or set a time for the analysis.  
**Note.** If an index is already listed in the *Current Activity* tab, either currently being processed or queued, then the index cannot be scheduled.
    - Click on **Save and Close** if you do not want to run the index at this stage. You can run or schedule analysis using this or any index directly from the Index page.
    - Click on the **Cancel** button to return to the index page, discarding all information about this index.
- If you have saved the index, it is now appended to the list of indexes on the *Indexing Overview* tab.



## Adding Multiple Indexes

To create a batch of indexes using the same analysis options, click on the **Add Multiple Indexes** link on the *Indexing Overview* page. The *Add Multiple Indexes* page is displayed.

**Note.** You cannot add multiple indexes that overlap with each other. You can create overlapping indexes, but not in a single operation. This is because overlapping indexes only make sense when they have different configurations, for example, a skim at the parent level, and more detailed analysis in the children.

1. Type a prefix to identify the batch of indexes in the *Index name prefix* text box. The names of individual indexes are constructed from the prefix and the name of the deepest folder at the start location.
2. Choose the Index Configuration to be used for this batch of indexes. This determines the indexing, analysis and calculated fields. Three default index configurations (see below) or custom configurations can be defined on the Index Configurations tab (see page 87):
  - **Default Skim**  
Information is collected about the folder structure and basic properties of files and folders, without analysis of file content.
  - **Default Skim and Duplicate**  
This adds basic analysis of the content of each file, limited to identifying Duplicate (identical copy) documents. Duplicate analysis generates data for duplication reporting and cleansing. This is the minimum requirement for duplication analysis.
  - **Default Skim, Duplicate and Textual (Analysis Pack license required)**  
This adds Textual (text content) analysis in addition to the previous levels of detail. Textual analysis includes File Format identification, Thematic analysis, Keyword and Text Content Pattern extraction, and File and Repository Property analysis. File Format analysis allows you to set up reports based upon the exact file format determined by inspecting file contents rather than filename extension. Thematic analysis consists of the extraction of file metadata to use as themes and/or summaries and to find similar files. It includes the keywords, description/comments and title fields from both Office style files and HTML web pages.
3. List the **Start Locations** of the content that you want to analyze using this batch of indexes.
  - Choose the location type, for example: File System.
  - Type the locations, entering one location per line.If the location is not present on the network map, it will be added, provided you have the necessary permissions. See your System Administrator if this is not the case.
4. Analysis normally proceeds with the credentials used by the Scheduler Service configured by the Systems Administrator (see the *Installation Guide* for details). In situations where these credentials are not sufficient to analyze one or more of the listed locations, indexing will fail unless you select appropriate credentials from the **Network Credentials** dropdown list.
5. Save the Indexes:
  - Click on **Save and Schedule** if you want to run the indexes immediately or set a time for the analysis.

**Note.** If an index is already listed in the *Current Activity* tab, either currently being processed or queued, then the index cannot be scheduled. If this occurs, Discovery Center will reject the batch of indexes.
  - Click on **Save and Close** if you do not want to run the indexes at this stage. The indexes are appended to the list on the *Indexing Overview* page. You can run or schedule analysis using these or any index directly from the *Indexing Overview* page.
  - Click on the **Cancel** button to return to the *Indexing Overview* page, discarding all information about this batch of indexes.





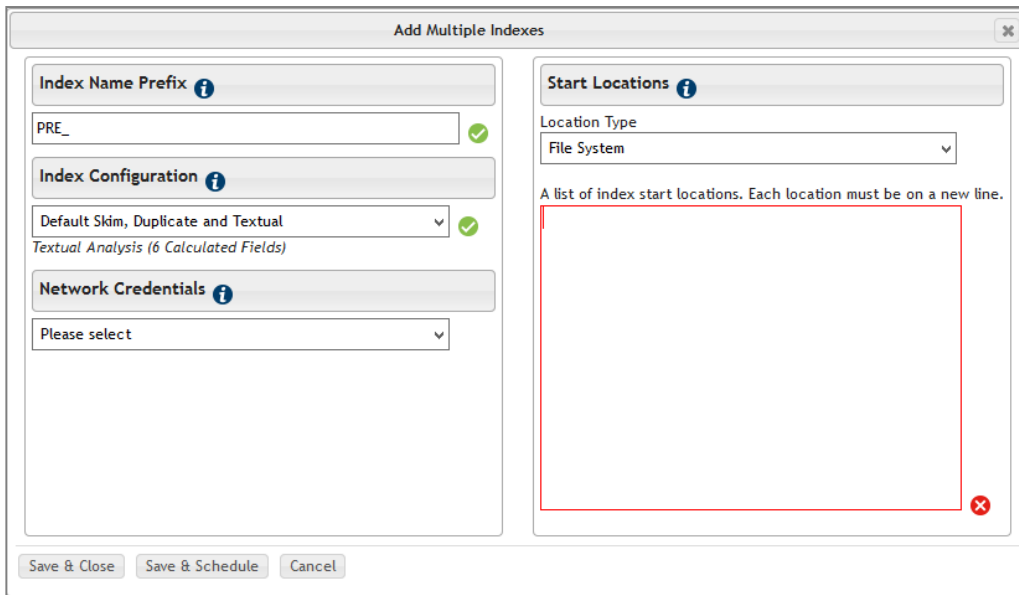


Figure 42 The Add Multiple Indexes dialog box

## Optimizing an Index

Depending on your choice of options and the size of the target file store, analysis can take a long time to complete. It is therefore important to select the right options for the intended purpose of the index and to avoid unnecessary and time-consuming analysis. There are three main purposes for an index:

- Analysis of file information for reporting purposes
- Cleansing of redundant and duplicate files
- Classification and migration

The following table lists the recommended settings for each of these tasks:

Table 13 Optimizing an index

Setting	Report/Audit	Cleansing	Migration
Duplicate	Y	Y	Y
Textual	Optional	Optional	Y
Thematic (1)	N	Optional	Y
Look inside zip files (2)	Optional	Optional	N
Duplicate	Y	Y	Y

(1) Thematic analysis extracts the key themes and summaries from the text of documents and is one part of textual analysis. Themes and summaries might be used to populate the Keywords and Summary properties of documents in SharePoint for example, and



themes are often used in Classification Workbench/Designer node search terms. If it is not required, thematic analysis can be turned off in the index configuration (see page 92).

(2) The expansion of Zip files can dramatically increase the quantity of information you will record during analysis and affect the duration of the detailed analysis phase. It also requires space in the server file cache folder to expand the entire content of the archives during analysis. Before choosing this option, (disabled by default) in an index configuration, it is advisable to perform an initial skim to assess the number and size of Zip files on your system.

## Optimizing Analysis Performance

Full details explaining how to improve and optimize index performance are provided in Appendix 1 on page 177. However, when configuring indexes consider the following factors for best performance:

- **Network Performance**  
Analysis requires that files are copied from their storage location to the Discovery Center file cache. For this reason, analysis performance is heavily dependent upon network quality and performance. Where analysis occurs across a Wide Area Network or where files are stored in an archive solution, performance can be significantly reduced.
- **First/All Matching Extraction Rules**  
For each additional extraction rule in use, analysis performance will be reduced.
- **Large Files**  
The larger a file is, the more time the analysis processes will take. Where an index includes very large files (such as those of more than 500 MB) it is not uncommon to observe analysis appear to 'hang' whilst those files are being moved across the network and analyzed. By default, Discovery Center excludes files over 500MB from textual analysis but not from Duplicates analysis. You can change these settings in the Index Configuration (see page 87). Certain large file types (such as data, log, or image files) can be safely excluded from analysis using exclusion rules in the index configuration.

## Scheduling an Index

The automatic running and re-running of indexes is controlled by a schedule. Scheduling may also be used to run indexes at a time when users will not be accessing the system or to avoid routine network events such as virus scans or backup operations.

To set up a schedule for an index, or to edit or remove one that has already been assigned:

1. Click on the **Indexes** tab.
2. On the *Indexing Overview* tab, select the index or indexes you want to schedule.
3. Click on the **Schedule Selected** link at the top of the page.
4. Choose **Schedule Indexes** from the dropdown list.
5. Choose from the following three options:
  - **Add to the end of the queue**  
Queue the index for running now.
  - **Schedule in the future**  
Specify a date, time and repeat frequency.
  - **Remove from schedule**  
Cancel the current schedule: the index will be assigned an *unscheduled* status.
6. If you selected the *Schedule in the future* option, use the *Future Scheduling* section to identify the date and time you want to run the index.
  - Click on the 'calendar' icon. Select the day when you want to run the index.
  - Select a time to run the index from the dropdown list box. Configure crawls to run at a time that is convenient for your users and at a frequency that reflects the rate at which information is changing.



7. Having set a date and time when the index is to be queued for analysis, use the *Recurring Scheduling* section if you want the index to be updated automatically:
  - Select **Keep this index up-to-date** if you want the analysis to be updated regularly.
  - Use the list box to select a daily, weekly, monthly or yearly period basis.
  - In the first box, enter the frequency of updates.
8. Click on the **Save** button to apply the schedule. Alternatively, click on **Cancel** to discard any changes to the current schedule.

**Note.** If an index is already listed in the *Current Activity* tab, either currently being processed or queued, then the index cannot be scheduled. If you are attempting to schedule multiple indexes, Discovery Center will reject the batch, listing up to five of the indexes currently being processed.



# Activity

The Activity page has two tabs:

- **Current Activity**  
Shows information about the current task and lists all tasks queued for analysis.
- **Activity History**  
Lists completed tasks from a specified time period.


## Current Activity

The *Current Activity* page provides information about the progress of queued tasks. The Discovery Center can only carry out one task at a time. If it is busy then any additional tasks that are due to be carried out at that time will be added to a queue and the first item in the queue will be started when Discovery Center becomes available.

The Current Activity page is divided into two sections: *Running Task* and *Queued Tasks*.

### Running Task

Shows information about the currently running task:

- **Name**  
Description of task and, for index processing, name of index.
- **Status**  
Current activity, for example, *Analyzing*. Click on link to display Task Status.
- **Progress**  
Percentage of task completed.
- **Running Time**  
Time elapsed since commencement of task.
- **Est. Time Remaining**  
Estimated time remaining to completion.
- **Actions**   
Click on the **Stop** button to stop the task.

**Note.** It may take some time for analysis to stop and this will not be an immediate process as analysis completes and closes. It may result in the task queue standing empty for a short while until all analysis tasks are resolved.



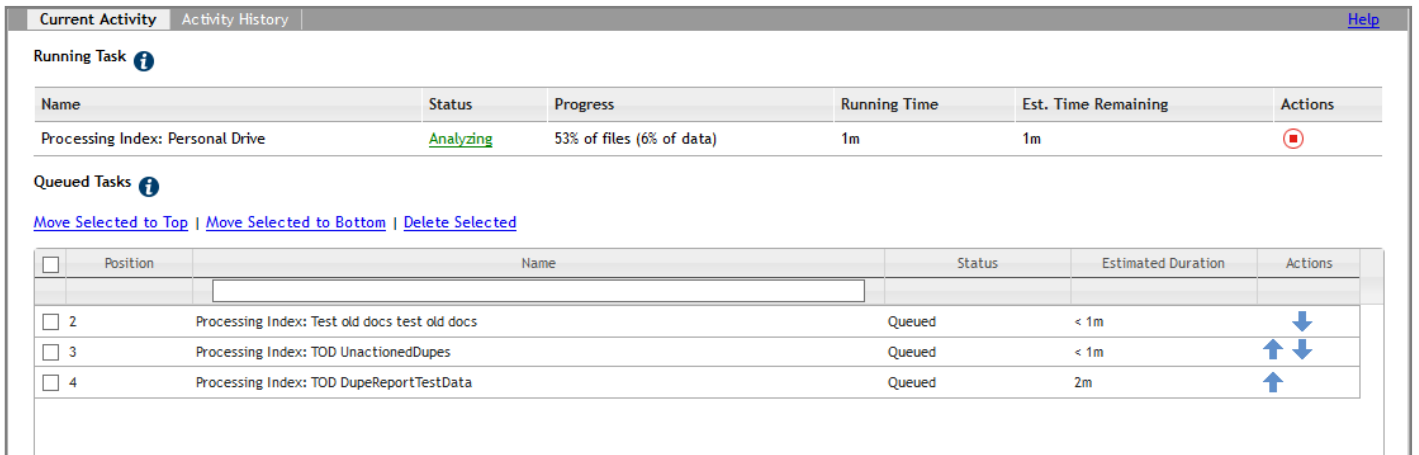
## Queued Tasks

Lists all indexes scheduled for analysis. The following links are displayed above the list:

- **Move Selected to Top**  
Promote the selected tasks to the front of the queue.
- **Move Selected to Bottom**  
Demote the selected tasks to the end of the queue.
- **Delete Selected**  
Remove selected tasks from the queue.

Queued tasks are listed under the following headings:

- **Position**  
Numerical position of the task in the queue.
- **Name**  
Description of task and, for index processing, name of index.
- **Status**  
Current activity, for example *Queued*. Click on the link to display a Task Status report (see page 103).
- **Estimated Duration**  
Estimation of task processing time.
- **Actions**  
Use the *Actions* buttons to move individual tasks up or down the queue.



The screenshot shows the 'Current Activity' page with two sections: 'Running Task' and 'Queued Tasks'. The 'Running Task' section displays a table with one row: 'Processing Index: Personal Drive' with status 'Analyzing', progress '53% of files (6% of data)', running time '1m', and estimated time remaining '1m'. The 'Queued Tasks' section has links for 'Move Selected to Top', 'Move Selected to Bottom', and 'Delete Selected'. Below these links is a table with columns for 'Position', 'Name', 'Status', 'Estimated Duration', and 'Actions'. The table contains three rows of queued tasks.

Current Activity		Activity History	Help		
Running Task ⓘ					
Name	Status	Progress	Running Time	Est. Time Remaining	Actions
Processing Index: Personal Drive	Analyzing	53% of files (6% of data)	1m	1m	⊞
Queued Tasks ⓘ					
<a href="#">Move Selected to Top</a>   <a href="#">Move Selected to Bottom</a>   <a href="#">Delete Selected</a>					
<input type="checkbox"/>	Position	Name	Status	Estimated Duration	Actions
<input type="checkbox"/>	2	Processing Index: Test old docs test old docs	Queued	< 1m	↓
<input type="checkbox"/>	3	Processing Index: TOD UnactionedDupes	Queued	< 1m	↑ ↓
<input type="checkbox"/>	4	Processing Index: TOD DupeReportTestData	Queued	2m	↑

Figure 43 Queued tasks on the Current Activity page



## Filtering

To filter the queued tasks by name:

1. Type a text string into the box at the top of the *Name* column.
2. Press the Return key.

The list is restricted to tasks that contain the specified text string(s). Click on the Reset Filter link at the top of the page to restore the full list of tasks.

## Sorting

Click on the column headers to reorder the listed tasks. Click on the column header a second time to reverse the sort order.

## Page Sizes

The *Queued Tasks* table presents tasks in pages according to the controls in the table footer. By default, each page lists up to 50 items although you can change this setting to 25 or 100 using the dropdown control. Use the other controls to browse through additional pages. Alternatively, type the page number you want to display.

Pages and task lists are automatically refreshed and sorted according to the chosen column header and ascending/descending order.

**Note.** *Database Cleanup* and *Reporting Database Processing* are priority tasks. When scheduled, these tasks are placed at the top of the queue. *Database Cleanup* occurs daily. Reporting Database Processing tasks will (by default) be placed on the queue after any activity (indexing or action) is completed, with a priority controlled by configuration in the *Reporting Settings* tab of the *Reporting and Actions* page (see page 173).

## Activity History

The *Activity History* page provides information about all tasks completed by Discovery Center within a specified time period. It documents the following activities:

- All Index processing tasks (one entry for each time the index is processed).
- All Reporting Action tasks (the deletion or migration of files initiated from the Actions menu or the chart context menu).
- All System administration tasks (database clean-up and optimization and reporting database processing).
- Metadata value import.
- Generating duplication report data.

By default, the page shows the 50 most recently completed tasks. To change the limit of shown tasks, alter the numeric value within the "# most recent activities" box. Inputting an empty value will remove the limit on shown tasks.

To define a time period, click on the  icons adjacent to the **From:** and **To end of:** boxes and then click on the **Filter** button to display tasks completed within the specified period (dates are inclusive: the list will show all tasks completed on the specified dates).

# most recent activities  From:  To end of:  [Filter](#) [Reset Filters](#)

Description	Type	Status	Duration	Completion Time
Process Reporting Database	Process Reporting Database	<a href="#">Complete</a>	1m	57 minutes ago
Processing Index: Local file share Analysis 1	Process Index	<a href="#">Complete</a>	< 1m	59 minutes ago
Process Reporting Database	Process Reporting Database	<a href="#">Complete</a>	1m	1 hours ago
Delete from 'test old docs' (based on 1 selection(s) on a Containers report)	Report Action	<a href="#">Complete</a>	< 1m	1 hours ago
Delete from 'test old docs' (selected files only)	Report Action	<a href="#">Complete</a>	< 1m	1 hours ago
Markup files with 1 field value(s) in 'test old docs' (selected files only)	Report Action	<a href="#">Complete</a>	< 1m	1 hours ago
Markup files with 1 field value(s) in 'test old docs' (based on 1 selection(s) on a Containers report)	Report Action	<a href="#">Complete</a>	< 1m	1 hours ago
Markup files with 1 field value(s) in 'test old docs' (based on 1 selection(s) on a Containers report)	Report Action	<a href="#">Complete</a>	< 1m	1 hours ago
Markup files with 1 field value(s) in 'test old docs' (based on 1 selection(s) on a Containers report)	Report Action	<a href="#">Complete</a>	< 1m	1 hours ago
Process Reporting Database	Process Reporting Database	<a href="#">Complete</a>	1m	1 hours ago
Processing Index: Local file share skim	Process Index	<a href="#">Complete (1 warnings)</a>	< 1m	1 hours ago
Database Cleanup	Database Cleanup	<a href="#">Complete</a>	< 1m	3 hours ago
Process Reporting Database	Process Reporting Database	<a href="#">Complete</a>	1m	2022/03/28 15:19
Processing Index: SP test docs	Process Index	<a href="#">Complete (1 warnings)</a>	< 1m	2022/03/20 14:18
Database Cleanup	Database Cleanup	<a href="#">Complete</a>	< 1m	2022/03/20 06:29
Process Reporting Database	Process Reporting Database	<a href="#">Complete</a>	1m	2022/03/17 11:28
Processing Index: SP test docs	Process Index	<a href="#">Complete (1 warnings)</a>	< 1m	2022/03/17 11:26

Page 1 of 1 50

Last updated: Thursday, 31 March 2022 14:32

Figure 44 Activity History page

Tasks are listed under the following headings:

- **Description**  
Type of activity and name of index, if relevant.
- **Status**  
Summary of task performance with basic error information linking to more detailed information in a Task Status report. This shows information about the activity, and for index processing tasks, statistics about the folders, files and analysis methods processed. If any problems were encountered during the activity, these are documented with a basic text description.
- **Duration**  
Task processing time.
- **Completion Time**  
Date and time when the task was completed.

### Filtering

You can filter the *Activity History* by Description (Index name/task) or Status:

1. Type a text string into the boxes at the top of the *Description* and/or *Status* columns.
2. Press the Return key.

The list is restricted to tasks that contain the specified text string(s). Click on the **Reset Filters** link at the top of the page to restore the full list of tasks.

### Sorting

Click on the column headers to reorder the listed tasks. Click on the column header a second time to reverse the sort order.



## Page Sizes

The *Activity History* table presents tasks in pages according to the controls in the table footer. By default, each page lists up to 50 items although you can change this setting to 25 or 100 using the dropdown control. Use the other controls to browse through additional pages. Alternatively, type the page number you want to display.

Pages and task lists are automatically refreshed and sorted according to the chosen column header and ascending/descending order.





## Task Status Reports

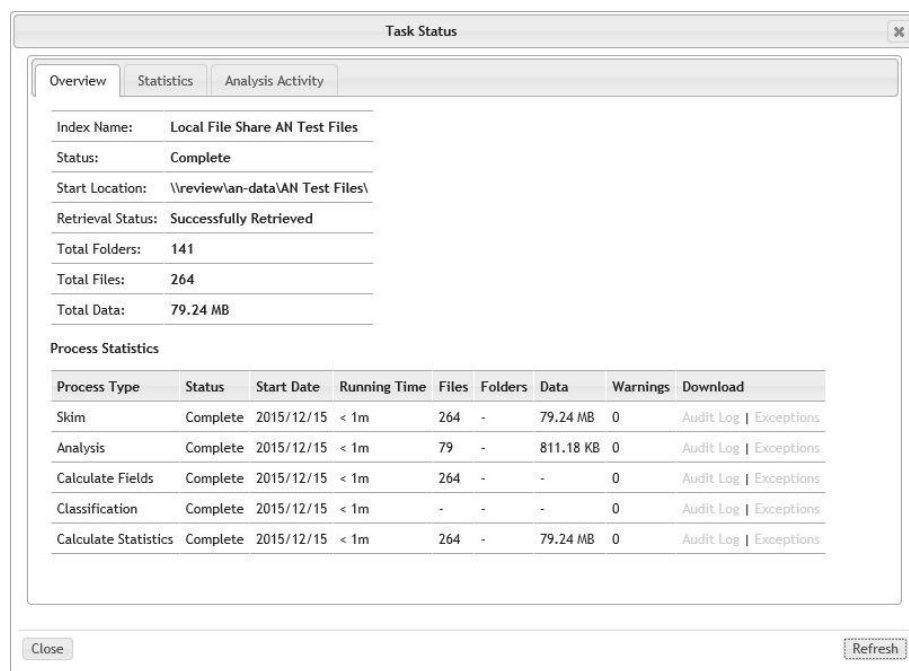
For any task listed on the *Current Activity* or *Activity History* pages, click on a **Status** link to view detailed information in a Task Status report. If any warnings have been issued during processing and files excluded from processing, the **Exceptions** link (in the *Download* column) is active. Click on this link to download a CSV file listing the files responsible for any warnings and more detailed error information. Any task involving file actions (see **'Actions' Task Status report**) such as migrate or delete also provides an audit trail: click on the **Audit Log** link to download a CSV file listing all files actioned.

### 'Processing Index' Task Status reports

For an indexing process, the Task Status report is divided into three tabbed pages:

- **Overview**  
Shows information about the index and statistics from the last occasion that the index was run. This includes a *Retrieval Status* message and statistics about the index and folders, files and analysis methods. If any warnings have been issued during processing and files excluded from processing, the appropriate **Exceptions** link (in the *Download* column) is active. Click on this link to download a CSV file listing the files responsible for any warnings and more detailed error information. Click on the **Audit Log** link to download a CSV file listing all files actioned.
- **Statistics**  
The Folder and File Statistics table shows detailed information about the number of files and folders processed during the task. The *Analysis Statistics* table lists information about the number of files analyzed for Metadata, File Format, Duplicate and Thematic content. This includes an *Analysis Status* message.
- **Analysis Activity**  
This tab provides information about an active index including the names of the files currently undergoing analysis and the duration of the task. This can be used to aid troubleshooting if an index appears to be running slowly. Click on the **Refresh** button to update the tab.

For further troubleshooting, click on the **Go to Application Log Files** page link.



The screenshot shows a 'Task Status' window with three tabs: Overview, Statistics, and Analysis Activity. The Overview tab is selected and displays the following information:

Index Name: Local File Share AN Test Files  
Status: Complete  
Start Location: \\review\an-data\AN Test Files\  
Retrieval Status: Successfully Retrieved  
Total Folders: 141  
Total Files: 264  
Total Data: 79.24 MB

Below this is a 'Process Statistics' table:

Process Type	Status	Start Date	Running Time	Files	Folders	Data	Warnings	Download
Skim	Complete	2015/12/15	< 1m	264	-	79.24 MB	0	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>
Analysis	Complete	2015/12/15	< 1m	79	-	811.18 KB	0	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>
Calculate Fields	Complete	2015/12/15	< 1m	264	-	-	0	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>
Classification	Complete	2015/12/15	< 1m	-	-	-	0	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>
Calculate Statistics	Complete	2015/12/15	< 1m	264	-	79.24 MB	0	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>

At the bottom of the window, there are 'Close' and 'Refresh' buttons.

Figure 45 An index processing Task Status report

## 'Actions' Task Status Report and Audit Trail

For any task involving file actions, the Task Status Report shows detailed information about the action including its type, source, options and any specific selections. The report also provides an audit log (CSV format) listing the actioned files and all actions including markup and delete/migrate operations. Click on the **Audit Log** link to download a CSV file listing the actioned files and information about metadata sets, the creation of shortcuts, and the name and location of files retained when processing duplicate files. Click on the **Exceptions** link to download a CSV file listing the files responsible for any warnings and more detailed error information.

Task Status

Action Type: Delete

Source: Area of Interest: Test Data

Action Options: Deleted files quarantined to location specified in Report Settings.

Actioned Selections: .exe

Report Name: File Extensions Report

Report Filter: -

Total Files: 15 of 24 files successfully actioned

Total Data: 564.83 MB of 742.87 MB

Created By: REVIEW\techauthor

Process Statistics

Process Type	Status	Start Date	Running Time	Files	Data	Warnings	Download
Snapshot	Complete	2014/06/02 13:18	1m	-	-	0	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>
Process Action	Complete	2014/06/02 13:19	< 1m	15	564.83 MB	9	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>

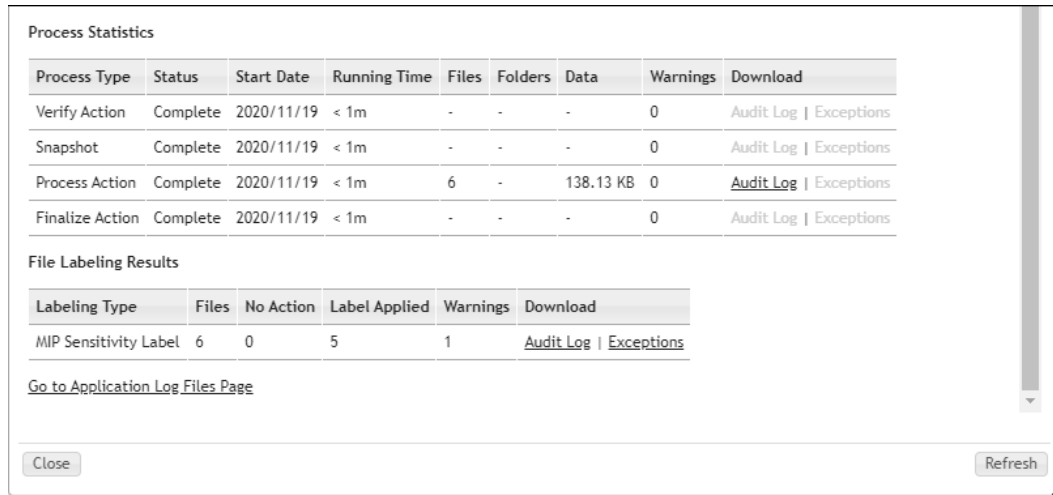
[Go to Application Log Files Page](#)

Close

Figure 46 An 'Actions' Task Status report

## File Labeling Status Report and Audit Trail

If the action included the application of MIP Sensitivity Labels to files then a separate status report is included providing details of the results of the MIP Sensitivity Label process for each file included in the action. The report table indicates how many of the files included in the action were excluded from the labeling process, how many were labeled successfully and how many incurred a failure. Full details of the successful and failed label applications for each file are included in the “Audit Log” and “Exceptions” downloads.



The screenshot shows a window with two main sections: 'Process Statistics' and 'File Labeling Results'. The 'Process Statistics' section contains a table with columns: Process Type, Status, Start Date, Running Time, Files, Folders, Data, Warnings, and Download. The 'File Labeling Results' section contains a table with columns: Labeling Type, Files, No Action, Label Applied, Warnings, and Download. Below the tables are links for 'Go to Application Log Files Page', 'Close', and 'Refresh'.

Process Type	Status	Start Date	Running Time	Files	Folders	Data	Warnings	Download
Verify Action	Complete	2020/11/19	< 1m	-	-	-	0	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>
Snapshot	Complete	2020/11/19	< 1m	-	-	-	0	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>
Process Action	Complete	2020/11/19	< 1m	6	-	138.13 KB	0	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>
Finalize Action	Complete	2020/11/19	< 1m	-	-	-	0	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>

Labeling Type	Files	No Action	Label Applied	Warnings	Download
MIP Sensitivity Label	6	0	5	1	<a href="#">Audit Log</a>   <a href="#">Exceptions</a>

[Go to Application Log Files Page](#)

Close Refresh

Figure 47 An ‘MIP Sensitivity File Labeling’ Status report



# Reporting and Actions

**Note.** Reporting performance depends upon SQL Server hardware specification, the number of files in the report and any other tasks being carried out by the Discovery Center host. The first report generated in a session will take additional time to establish a SQL Server Reporting session and reports across millions of files may take several minutes to produce.

Reports are compiled and stored in a reporting database, managed by the AN administrator. The database must be processed in order to reflect the most current system state. If a report is out of date or stale, the Discovery Center displays a status message explaining the possible causes: the index may have been rerun, files may have been migrated or deleted, there may have been changes to metadata definitions or an import of metadata, or changes to AOI definitions.

The *Reporting and Actions* page has six tabs:

- **Reporting Overview** (*Information Managers only*)  
Define and investigate (for reporting, cleansing or migration purposes) one or more locations within the Network Map.
- **Saved Views** (*Information Managers and Reviewers*)  
Create new reports and manage saved views.
- **Actions** (*Information Managers only*)  
View details of Actions applied across the Network Map. Create or edit schedules for re-running Actions.
- **Work Packages** (*Information Managers and Reviewers*)  
List Work Packages involving you as either an assigned reviewer or originator/owner.
- **Report Viewer** (*Information Managers only*)  
Report display area.
- **Mapping Rules** (*AN Administrators only*)  
Create metadata mapping sets for file migration, including rules for updating metadata, and create a set of substitute text strings to replace illegal characters in file and folder names.
- **Reporting Settings** (*AN Administrators only*)  
Configure the priority of the reporting database processing task, set archive location and file list size.

Reports enable decisions to be made about removing or deleting unwanted files, or migrating files based on metadata characteristics or classification results. Files can be deleted or migrated directly from reports (see page 151). You can also export tabulated data for your own reports and presentations.



# Reporting Overview

An Area of Interest (AOI) defines one or more locations within the Network Map identified by an Information Manager as having a business-relevant relationship. For example, files and folders relevant to, or owned by a finance team, represent their AOI. The definition of AOIs allows horizontal reporting (an Index, on the other hand, represents a vertical analysis of a folder, its subfolders and files) and the detailed exploration of related information.

The Reporting Overview tab is divided into two sections. There are two tabs on the left side:

- **Areas of Interest**  
Lists previously defined areas of interest. The tab has the following links:
  - **Add**  
Create a new Area of Interest (AOI) - see Adding a New Area of Interest.
  - **Delete Selected**  
Delete all selected AOIs (selected using the check boxes at the start of each row).
  - **Reset Filters**  
Remove any filter applied to the Name column and restore the full list of AOIs.
- **Locations**  
Shows the network map and allows you to select a specific folder or network location for reporting purposes.

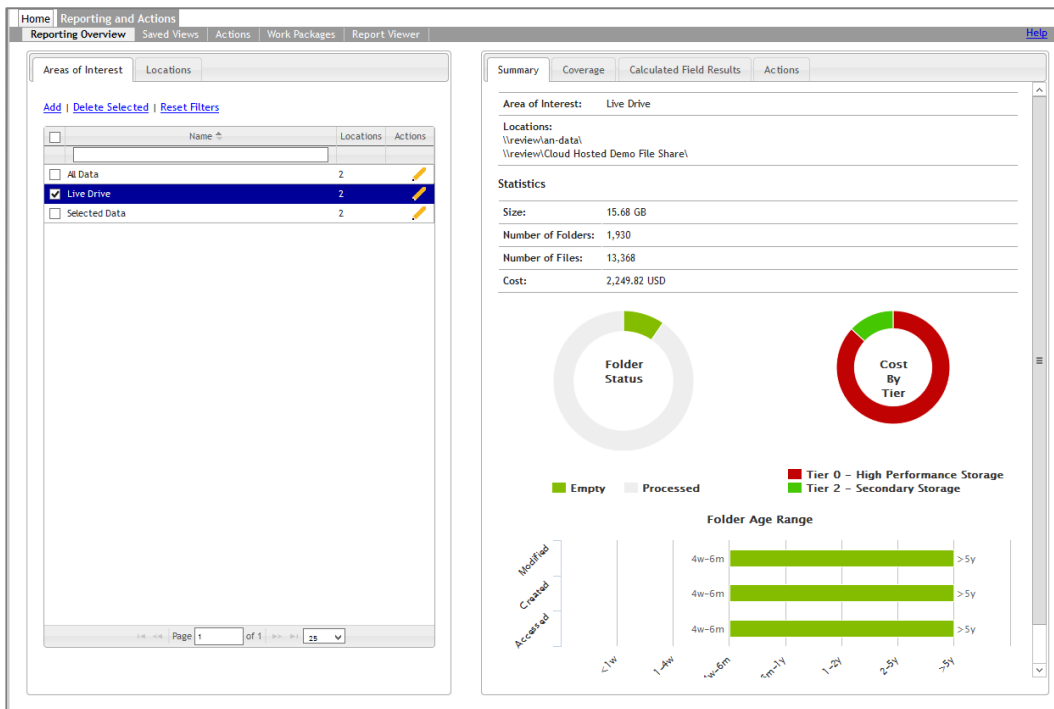


Figure 48 The Reporting Overview tab



## Adding a New AOI

To add a new Area of Interest (AOI):

1. On the *Areas of Interest* tab, click on the **Add** link.
2. Enter a *Name* for the AOI. Optionally, enter a text *Description*.
3. In the *Locations* box, browse the Network Map and select the folders to include in the AOI. As you do so, the folders are listed in the *Selected Locations* list.
4. Click on the **Save** button.

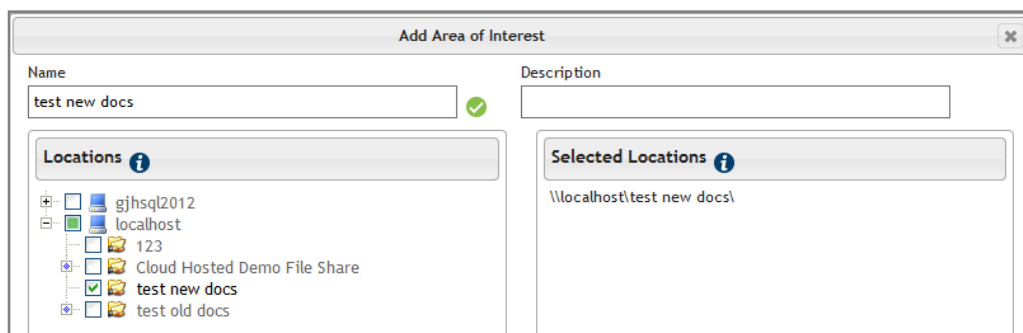


Figure 49 Adding an Area of Interest

More detailed information about the selected Area of Interest or location is shown on the right of the Reporting Overview tab on four tabbed pages:

- Summary
- Coverage
- Calculated Fields Results
- Actions

### Summary

Displays more detailed information about a selected Area of Interest or location: its name, type and retrieval status. In addition, the *Statistics* section shows the size and number of folders present in the selected location or Area of Interest. Two graphs show:

- **Folder Status**  
A pie chart illustrating the proportion of folders that are:
  - **Processed** (grey)  
Folders that have been skimmed
  - **Inaccessible** (red)  
Folders which cannot be accessed
  - **Empty** (purple)  
Folders with no content
  - **Extracted Archive folders** (yellow)  
ZIP files (only listed when *Include zip file content* option has been enabled in Index Configuration)
- **Cost by Tier**  
A pie chart showing tier assignment and cost within the chosen location.
- **Folder Age Range**  
A bar chart illustrating the span of file ages, as determined by modified, created and accessed dates. For example, the following chart shows that some folders have been accessed recently while others have been inactive for more than 5 years.



Statistics

Size:	15.68 GB
Number of Folders:	1,930
Number of Files:	13,368
Cost:	2,249.82 USD

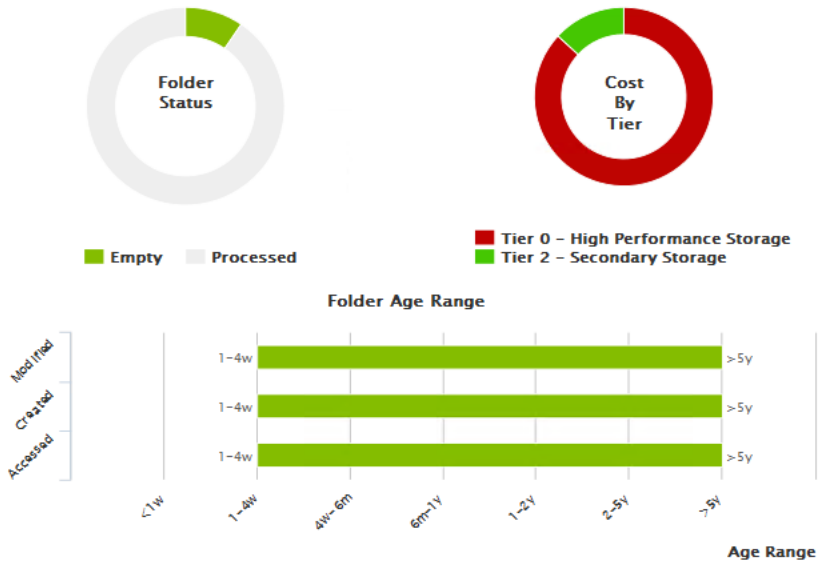


Figure 50 Summary tab

Coverage

This tab is divided into two sections:

- **Textual Analysis Coverage**

A pie chart illustrating the extent of textual analysis in the location or Area of Interest. The green segment illustrates the proportion of files in the selected locations where textual analysis has been attempted. The remainder (labeled *Analysis Not Attempted*), have been discovered (skimmed) and may have been subjected to duplicate analysis but have not had full textual analysis attempted. The outer ring of the green segment shows the proportion of files where textual analysis has been successful (green), excluded (yellow) or where errors have occurred (red).

**Note.** Files may have been excluded because of the index configuration, for example, constraints on file size or extension. Information about any textual analysis errors can be found in the activity log for the particular index.

- **Calculated Field Coverage**

Lists the calculated fields applied to the content, color-coded to show the most commonly applied fields. A calculated field shown in green text is applied across the whole location or Area of Interest. Fields displayed in orange text have only been applied partially across the location/Area of Interest.



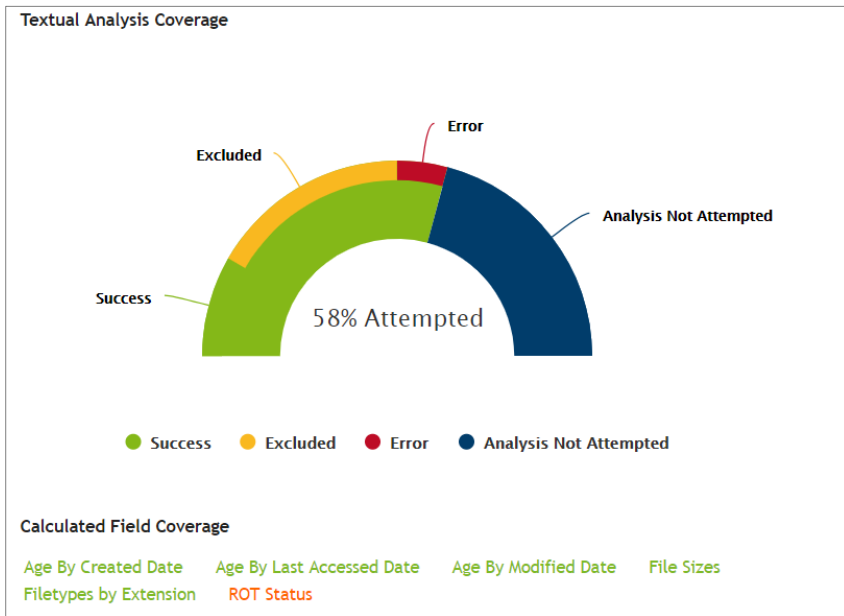


Figure 51 Coverage tab

### Calculated Field Results

This tab displays a bar chart showing the number of matching files for all discovered calculated fields. Hover over an individual bar to see the actual number of matched files. Click on the button in the top right corner to view a full-screen version of the chart.

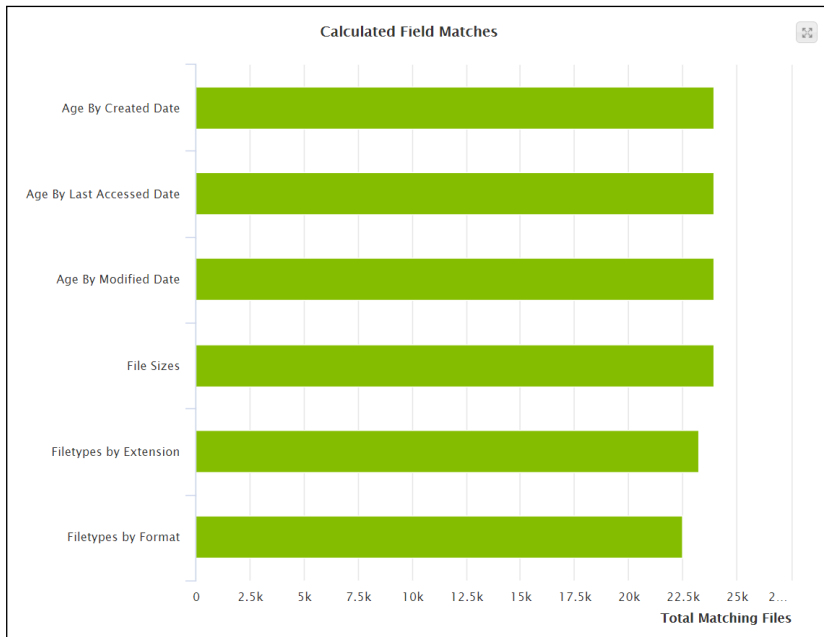


Figure 52 Calculated Field Results tab





## Actions

Use the “Actions” tab to create a report for the selected Area of Interest or network location.

### Define View

Use the **Report Type** dropdown list to select a view type. Click on the **Generate** link to generate the view. Discovery Center displays the *Report Viewer* tab and generates the relevant view (if the data is available).

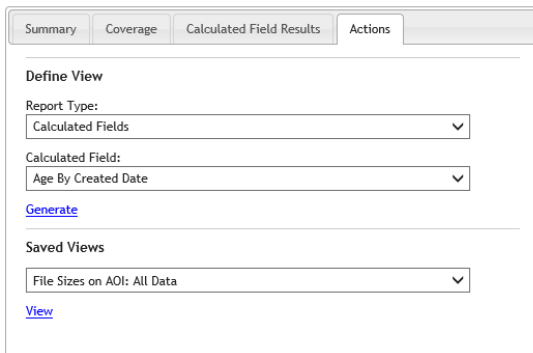
The following view types are available:

- **Calculated Fields**  
An additional dropdown box is displayed. Choose the *Calculated Field* to be displayed from those listed.
- Containers
- **Content and File Duplicates**  
An additional dropdown box is displayed to allow master selection strategy to be selected.
- **File Duplicates**  
An additional dropdown box is displayed to allow master selection strategy to be selected.
- File Extensions
- Files by Created Date
- Files by Last Accessed Date
- Files by Modified Date
- Files By Owner

### Saved Views

Lists saved views for the chosen locations and report type selected in *Define View*.

Click on the **View** link to display a selected report.



The screenshot shows a web interface with a tabbed menu at the top containing 'Summary', 'Coverage', 'Calculated Field Results', and 'Actions'. The 'Actions' tab is active. Below the tabs is a 'Define View' section with two dropdown menus: 'Report Type' set to 'Calculated Fields' and 'Calculated Field' set to 'Age By Created Date'. A blue 'Generate' link is positioned below these menus. Underneath is a 'Saved Views' section with a dropdown menu showing 'File Sizes on AOI: All Data'. A blue 'View' link is located below the 'Saved Views' dropdown.

Figure 53 Actions tab

## Saved Views

The *Saved Views* tab displays a tabulated list of saved report views. The list can be ordered alphabetically or filtered by Name, AOI/Location, Report Type or Author. To apply a filter, type it into the text box in the appropriate header column. If the currently active filters exclude all saved views click on the **Reset Filters** link to display an unfiltered list.

**Note.** Ensure that the reporting database has been recently processed by your AN Administrator. If the reporting database has not been processed, a saved view may not truly reflect the state of current data. If you commit any actions when the reporting database is out of date, no data will be lost but your actions will be based upon the contents of the reporting database at the time.

<input type="checkbox"/>	View Name ↑	AOI / Location	Report Type	Author	Quick Load	Assigned	Actions
<input type="checkbox"/>	A Markup Field on \\localhost\Cloud Hosted Demo File Share\RHI\...	\\localhost\Cloud Hosted Demo File Share\RHI\...	Calculated Fields	REVIEW\activenav			
<input type="checkbox"/>	Containers by Count	\\localhost\test old docs\	Containers	REVIEW\activenav	✓		
<input type="checkbox"/>	Containers by Count 2	\\localhost\test old docs\	Containers	REVIEW\activenav			
<input type="checkbox"/>	Containers by Count on AOI: Live Drive with a one	Live Drive	Containers	REVIEW\InfoManager	✓		
<input type="checkbox"/>	Containers by Count on AOI: TOD by another name	TOD	Containers	REVIEW\activenav			
<input type="checkbox"/>	Containers Heat by PErcent	\\localhost\test old docs\	Containers	REVIEW\activenav			
<input type="checkbox"/>	Document Type on AOI: TOD by another name	TOD	Calculated Fields	REVIEW\activenav			
<input type="checkbox"/>	File Extensions on P Drive\TSmith\'	\\localhost\Cloud Hosted Demo File Share\RHI\...	File Extensions	REVIEW\activenav	✓		
<input type="checkbox"/>	Generic IM Policy ROT on AOI: TOD	TOD	Calculated Fields	REVIEW\InfoManager	✓		
<input type="checkbox"/>	My new view	\\gjsq(2012)\test old docs\	Calculated Fields	REVIEW\activenav			
<input type="checkbox"/>	My new view	\\localhost\	Calculated Fields	REVIEW\activenav			
<input type="checkbox"/>	My new view	\\localhost\	File Extensions	REVIEW\activenav			
<input type="checkbox"/>	My new view	\\gjsq(2012\	File Extensions	REVIEW\activenav			




Figure 54 Saved Views tab

The *Saved Views* tab has the following links:




- **Update Selected**
  - **Apply Quick Load**  
By enabling Quick Load, the view will be cached the next time the Reporting Database is processed. Without Quick Load enabled, a view may take some time to load.
  - **Assign Users or Groups**  
Saved views can only be accessed by Information Managers or Reviewers assigned by the view owner. Use the *Assign Users or Groups* link to provide access to the selected view(s) for specific users or groups with Reviewer or Information Manager privileges. Users and groups are prefixed with either '[U]' or '[G]' respectively to aid identification.
  - **Delete Selected**  
Delete one or more selected views (using the check boxes).
  - **Define Work Package**  
Request a review of the selected saved views in the Define Work Package dialog box (see page 115).
  - **Reset Filters**  
Display an unfiltered list. The list of saved views can be filtered by View Name, AOI/Location, Report Type or Author. To apply a filter, type it into the text box in the appropriate header column.



By default, the page only shows the views you have saved. If you want to see all views saved by all users, select the **Show all Saved Views** check box. Users or groups that only have *Reviewer* role rights can only see saved views explicitly assigned to them by an Information Manager. Saved views are listed under the following column headings:

- **View Name**  
Name given to the view when it was created. The default naming structure consists of the report type and AOI/location. You can rename views using the link in the Actions column or by double clicking on the view row. View names can be the same providing the report type or location is different.
- **Aoi/Location**  
Area of Interest or network location targeted by view.
- **Report Type**  
Type of report selected in the *Define View* dialog box.
- **Author**  
Username of view creator.
- **Quick Load**  
 Indicates whether cached results are available for this view.
- **Assigned**  
 Indicates whether the view has been assigned to one or more reviewers.
- **Actions**
  -  Edit view settings in the *Display View Settings* dialog box:
    - Edit the **View Name** as required.
    - Select **Enable Quick Load** to automate caching of data for the view when the reporting database is updated to allow faster display.
    - Users and groups with the appropriate roles are shown in the **Available Users and Groups** box prefixed with either '[U]' or '[G]' respectively to aid identification.
    - Reviewers can only access views assigned to them. Use the **Available/Selected Users and Groups** boxes to grant access to specific Reviewers.

   Open the selected view in the *Report Viewer* tab:

-  This view has been saved in the database which enables reports on large numbers of documents to be displayed faster. Views are automatically cached after they have been loaded, and the cached data is discarded automatically when the reporting database is reprocessed. If a View is based on an Area of Interest, and the locations in the Area of Interest are changed, then new results will be calculated next time the report is viewed.
-  The report may take longer to load because it has not been cached.
-  This view has been saved in the database, but no results were found.

## Defining a Work Package

A Work Package is created from a Saved View and has a description informing the Reviewer or group of Reviewers of the actions required, for example, to identify files to be deleted.

- When creating a Work Package, the Reviewer can be either a single Reviewer or a group of Reviewers. A group of reviewers is a Windows group (Active Directory or local Windows group on the Discovery Center server).
- If a group of Reviewers is selected, only one Work Package is created and assigned to the entire group. A member of the Reviewer group will activate the Work Package, which will 'lock' the Work Package preventing another Reviewer from activating it.

- Each Work Package has a deadline for the work to be completed. The Information Manager is notified when a Work Package is Activated or Completed by an assigned user. The Reviewer is notified when a Work Package is assigned or marked completed if assigned to a single user.

To create a new Work Package:

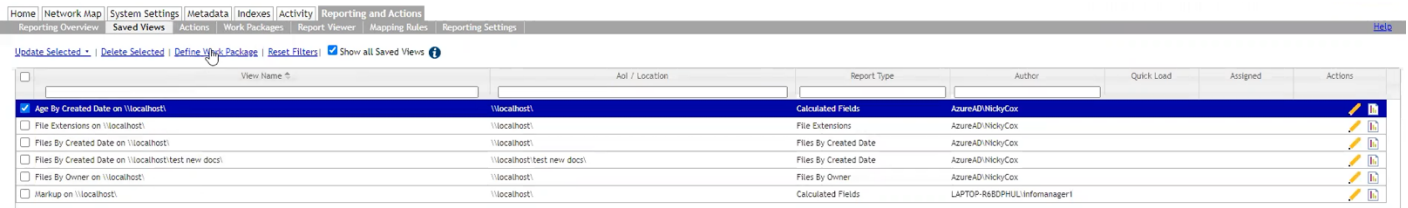


Figure 55 Create a Work Package

1. Open the **Saved Views** tab on the Reporting and Actions page.
2. Select the view or views that require input from the Reviewer.
3. Click on the **Define Work Package** link. The Define Work Package dialog box is displayed.
4. Enter the following information:
  - **Title:** Enter a name for the Work Package. By default, the Work Package is given the name of the first selected view.
  - **Description:** Type a message for the Reviewer, explaining the purpose of the review.
  - **Markup Field:** Choose a markup field for the reviewer to respond.
  - **Reviewer:** Select the [U] username or [G] group name of the Reviewer(s) to be assigned this Work Package.
  - **Deadline Date:** Specify the required completion date.
5. Click on **Save**.

The Work Package is added to the list displayed on the *Reporting and Actions - Work Packages* tab. Notifications regarding its status are displayed for the Reviewer and originating Information Manager.

## Actions

The page lists all Actions that have previously been applied from the Chart, Data Table and Container List tabs of a Report View. This list can be used to select specific actions to be re-run or to be scheduled for periodic action. This allows Information Managers to automate the cleansing of content according to information management policies.

Actions are listed with their properties under the following column headings:

- **Action Description**  
System-generated description of the Action and the targeted location.
- **Status**  
Whether Action was completed successfully when last run.
- **Owner**  
Username of Action initiator.
- **Scheduling**  
Scheduling status of Action.
- **Last completed**  
Date and time the Action was last applied.
- **Actions**
  - View details about the Action such as the selections and options associated with it.
  - Create edit or remove a schedule.



## Filtering

To filter the list of Actions:

1. Type the text string into the boxes at the top of the *Action Description*, *Status*, *Owner* and/or *Scheduling* columns.
2. Press the Return key.

The Actions list is restricted to items that contain the specified text string(s). Click on the **Reset Filters** link at the top of the page to restore the full list.


## Sorting

Click on the column headers to reorder the list. Click on the column header a second time to reverse the sort order.

## Scheduling an Action

The automatic re-running of previously enacted Actions can be controlled by a schedule. Scheduling may also be used to run Actions at a time when users will not be accessing the system or to avoid routine network events such as virus scans or backup operations.

To set up a schedule or to edit or remove one that has already been assigned:

1. For the required Action, click on the **Reschedule Action** icon  in the Actions column.
2. Choose from the following three options:
  - **Add to the end of the queue**  
Queue for running immediately.
  - **Schedule in the future**  
Specify a date, time and repeat frequency.
  - **Remove from schedule**  
Cancel the current schedule: the Action will be assigned an *Unscheduled* status.
3. If you selected the **Schedule in the future** option, use the *Future Scheduling* section to identify the date and time you want to run the Action.
  - Click on the 'calendar' icon. Select the day when you want to run the Action.
  - Select a time to run the Action from the dropdown list box. Configure an Action to run at a time that is convenient for your users and at a frequency that reflects the rate at which information is changing.
4. If you want to run the Action on a regular schedule, select **Run the Action periodically?**. Then, in the *Recurring Scheduling* section:
  - Use the list box to select a daily, weekly, monthly or yearly period basis.
  - In the first box, enter the frequency of updates. The Action will be carried out at the time specified in the *Future Scheduling* section.



Click on the **Save** button to apply the schedule. Alternatively, click on **Cancel** to discard any changes to the current schedule.









Home   Reporting and Actions						
Reporting Overview		Saved Views	Actions	Work Packages	Report Viewer	Help
Reset Filters						
Action Description	Status	Owner	Scheduling	Last Completed	Actions	
Markup files with 1 field value(s) in 'Live Drive' Aol (based on 1 selection(s) on a File Extensions report)	Complete	REVIEW\techauthor	Unscheduled	2017/01/02 12:06	 	
Delete from 'Live Drive' Aol (based on 1 selection(s) on a File Extensions report)	Complete	REVIEW\techauthor	Unscheduled	2017/01/02 12:05	 	
Markup files with 1 field value(s) in 'Cloud Hosted Demo File Share' (based on 1 selection(s) on a Filetypes t	Complete	REVIEW\activenav	Unscheduled	2016/12/12 12:55	 	
Delete from 'All Data' Aol (based on 1 selection(s) on a File Extensions report)	Error	REVIEW\techauthor	Unscheduled	-	 	
Delete from 'Live Drive' Aol (based on 1 selection(s) on a Age By Created Date field report)	Running	REVIEW\techauthor	-	-		

Figure 56      Actions tab



## Work Packages

Work Packages are created and managed by **Information Managers** and reviewed by an assigned **Reviewer** or a member of an assigned **Reviewer Group**.

A Work Package is created from a Saved View and has a description guiding the Reviewer on the actions required, for example: identifying files to be deleted.

Each Work Package has a deadline for the work to be completed. Notifications inform the Information Manager and Reviewer about the status of the Work Package from its inception to completion.

### Information Manager

**Note.** By default, the Work Packages tab for an **Information Manager** only displays the Work Packages that they created and that are in progress (Approved Work Packages are filtered out). Selecting **Show All Work Packages** will display all the Work Packages in the system, except for any deleted Work Packages.

An Information Manager is responsible for interrogating a network with Discovery Center, for example, to identify policy violations and remove ROT. In the process, the Information Manager generates reports highlighting files and folders for subsequent actions.

In many cases, the Information Manager requires the input of the business users and file owners before proceeding with actions such as archiving, deletion, or migration. To aid this process, the Information Manager can create a **Work Package** for a Reviewer or Reviewer Group responsible for the files or business area.

Title	Assigned User	Active User	Owner	Status	Deadline	Completed	Total Results	Report Status	Actions
SV1 - Rev1	LAPTOP-D4PA7NF5:Reviewer1		LAPTOP-D4PA7NF5:Infomanager	Assigned	Due in 6 day(s)	-	14		
sv2 - rev2	LAPTOP-D4PA7NF5:Reviewer2	LAPTOP-D4PA7NF5:Reviewer2	LAPTOP-D4PA7NF5:Infomanager	Active	2021/03/29	-	14		
sv2	LAPTOP-D4PA7NF5:Reviewer2		LAPTOP-D4PA7NF5:Infomanager	Assigned	2021/04/21	-	14		
SV1	LAPTOP-D4PA7NF5:REVIEWERS	LAPTOP-D4PA7NF5:Infomanager	LAPTOP-D4PA7NF5:Infomanager	Active	Tomorrow	-	14		

Figure 57 Work Packages Information Manager View

### Reviewers and Reviewer Groups

**Note.** By default, the Work Packages tab for a **Reviewer** only displays the items **assigned to them** (either directly or as a member of an assigned **Reviewer Group**). Selecting **Show All Work Packages** will display all the Completed and Approved Work Packages assigned to them. The Work Packages view never displays deleted Work Packages.

A Reviewer can be assigned directly or be part of a Reviewer Group that is assigned to a Work Package. Once assigned to a Work Package, it is the Reviewer's responsibility to activate and progress the Work Package.

Title	Assigned User	Active User	Owner	Status	Deadline	Completed	Total Results	Report Status	Actions
Files By Last Accessed Date on \\laptop-r8bdphul	LAPTOP-R8BDPHUL:Reviewer2	LAPTOP-R8BDPHUL:Reviewer2	LAPTOP-R8BDPHUL:Infomanager1	Active	2021/03/25	-	13,036		
File Extensions on \\laptop-r8bdphul	LAPTOP-R8BDPHUL:Reviewer2		LAPTOP-R8BDPHUL:Infomanager1	Assigned	2021/03/25	-	13,036		
Files By Created Date on \\laptop-r8bdphul	LAPTOP-R8BDPHUL:REVIEWERS		LAPTOP-R8BDPHUL:Infomanager1	Assigned	2021/03/26	-	13,036		

Figure 58 Work Packages Reviewer View






The Work Packages tab lists Work Packages under the following column headings:

- **Title**  
Name given to the package when it was created. The default naming structure consists of the report type and AOI/location.
- **Assigned User**  
Reviewer or Reviewer Group assigned to the Work Package.
- **Active User**  
Reviewer or member of Reviewer Group responsible for progressing the Work Package.
- **Owner**  
Information Manager responsible for generating the Work Package.
- **Status**  
Indicates the progress of the Work Package:
  - **Assigned:** An Information Manager has created the Work Package and assigned it to a Reviewer or Reviewer Group.
  - **Active:** The assigned Reviewer (either directly or as a member of the assigned Reviewer Group) has activated the Work Package. When the work is complete, the Reviewer clicks on the **Mark Review Completed** button in Report Viewer to advance the Work Package status to Complete.
  - **Complete:** The Reviewer has marked the Work Package as complete. If you are the owner of the Work Package, click on the link and then click on Approve Work Package to advance its status to Approved.
  - **Approved:** The Work Package owner has reviewed the work carried out and approved the Work Package.
  - **Deleted:** The Information Manager has deleted the Work Package. When a Work Package is deleted, it is removed from any Work Package view.

**Note.** If no actions have been performed on the Work Package, it is deleted from the database. If actions were performed on the Work Package, the Work Package's status is set to Deleted (to maintain an audit trail) but is no longer visible in the Work Package views.

Click on the links to view information about the Work Package, including its status and instructions from the originating Information Manager. The Reviewer and owner can choose to receive web and email notifications of these steps - see page 24.

- **Deadline**  
Completion date requested by the owner.
- **Completed**  
Date completed by the assigned Reviewer.
- **Total Results**  
Number of results requiring Reviewer input.
- **Report Status**
  -  indicates that a cached report related to this Work Package is available but contains no results.
  -  indicates that a cached report related to this Work Package is available for viewing.
  -  is displayed if a cached report is not available: it may therefore take a while to generate the view.
- **Actions**

Click on  to launch the Work Package. The saved view is displayed in the *Report Viewer* tab.

Click on  to activate the Work Package; a Report is generated and displayed.

Click on  to deactivate the Work Package; the Active User is removed, and the Work Package returns to assigned status.

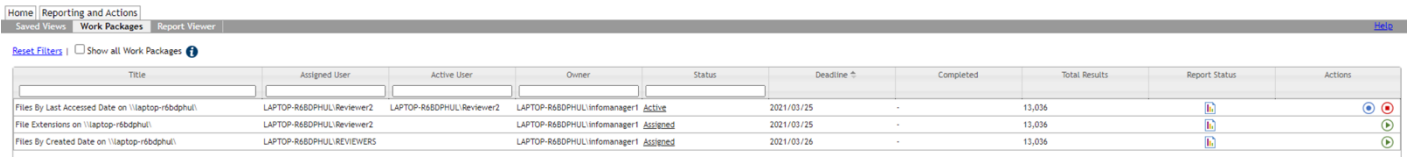


## Activating

A Work Package can be activated in two ways by the assigned Reviewer (or member of the assigned Reviewer Group), either by using the activate action icon  or via the Review Package Status dialog. To activate a Work Package:

### Activate via activate action icon

1. Select the **Work Packages** tab.
2. Locate the Work Package to be activated in the list. Click on the activate action icon  in the Actions column of the grid.












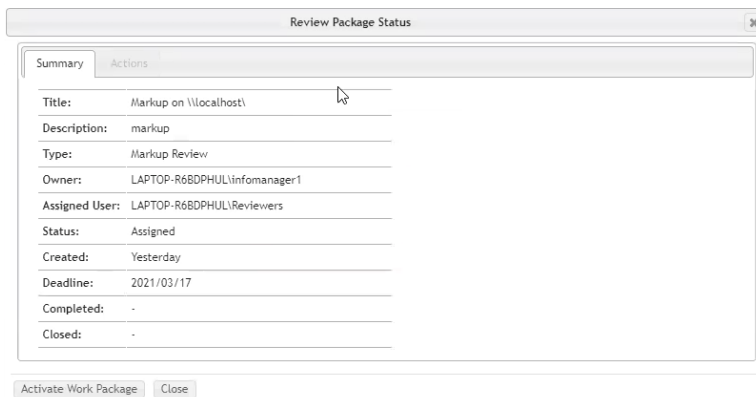
Title	Assigned User	Active User	Owner	Status	Deadline	Completed	Total Results	Report Status	Actions
Files By Last Accessed Date on \\laptop-r6bdphul\	LAPTOP-R6BDPHUL\Reviewer2	LAPTOP-R6BDPHUL\Reviewer2	LAPTOP-R6BDPHUL\infomanager1	Active	2021/03/25	-	13,036		 
File Extensions on \\laptop-r6bdphul\	LAPTOP-R6BDPHUL\Reviewer2	LAPTOP-R6BDPHUL\Reviewer2	LAPTOP-R6BDPHUL\infomanager1	Assigned	2021/03/25	-	13,036		 
Files By Created Date on \\laptop-r6bdphul\	LAPTOP-R6BDPHUL\REVIEWERS	LAPTOP-R6BDPHUL\REVIEWERS	LAPTOP-R6BDPHUL\infomanager1	Assigned	2021/03/26	-	13,036		 

Figure 59 Activate Work Package via activate icon

### Activate via the Review Status dialog

1. Select the **Work Packages** tab, then click on the status of the Work Package.
2. The Review Package Status dialog will appear. Click the **Activate Work Package** button.



Review Package Status

Summary Actions

Title: Markup on \\localhost\

Description: markup

Type: Markup Review

Owner: LAPTOP-R6BDPHUL\infomanager1

Assigned User: LAPTOP-R6BDPHUL\Reviewers

Status: Assigned

Created: Yesterday

Deadline: 2021/03/17

Completed: -

Closed: -

Activate Work Package Close

Figure 60 Activate Work Package via Status

A Report View will now be generated, and the Work Package's status will change to Active.

## Deactivating

A Work Package can be deactivated in three ways by the Active Reviewer - Using the deactivate action icon , Deactivating via the Review Package Status dialog, and Deactivating via the Work Package Report. To deactivate a Work Package:

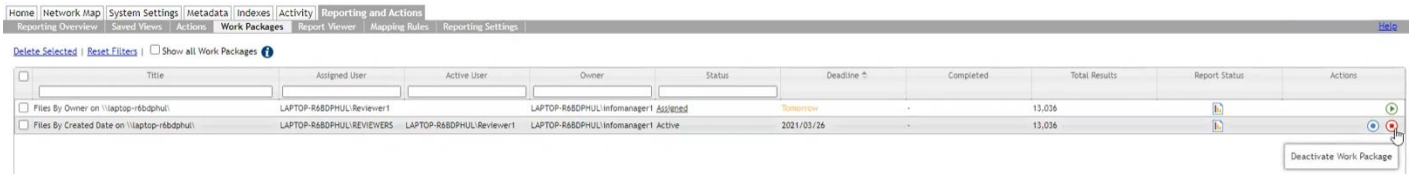


Figure 61 Deactivating a Work Package

### Deactivate via deactivate action icon

1. Select the **Work Packages** tab.
2. Locate the Work Package to be deactivated and click on the deactivate action icon  in the Actions column of the grid.

### Deactivate via Review Package Status dialog

1. Select the **Work Packages** tab.
2. Locate the Work Package to be deactivated and click on the status of the Work Package; the Review Package Status dialog will appear; click the **Deactivate** button.

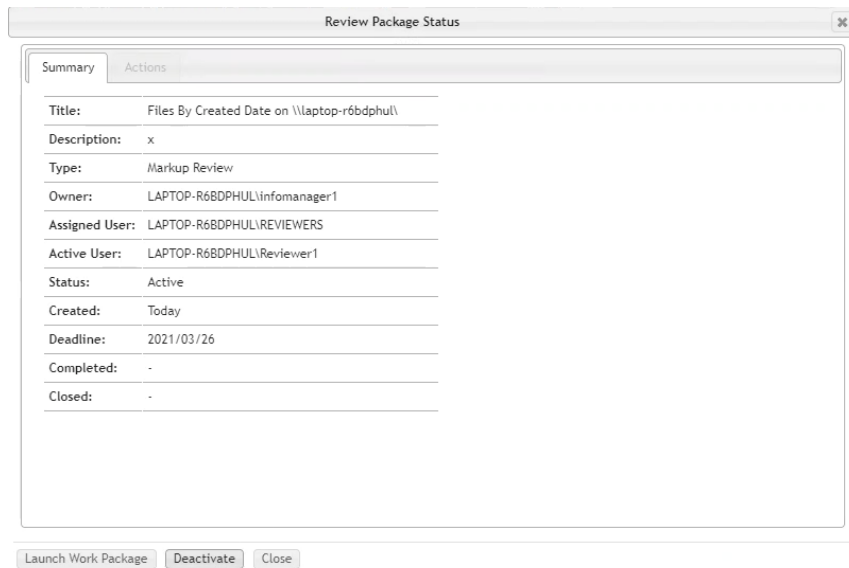


Figure 62 Deactivating a Work Package via Review Package Status

## Deactivate via the Report Viewer

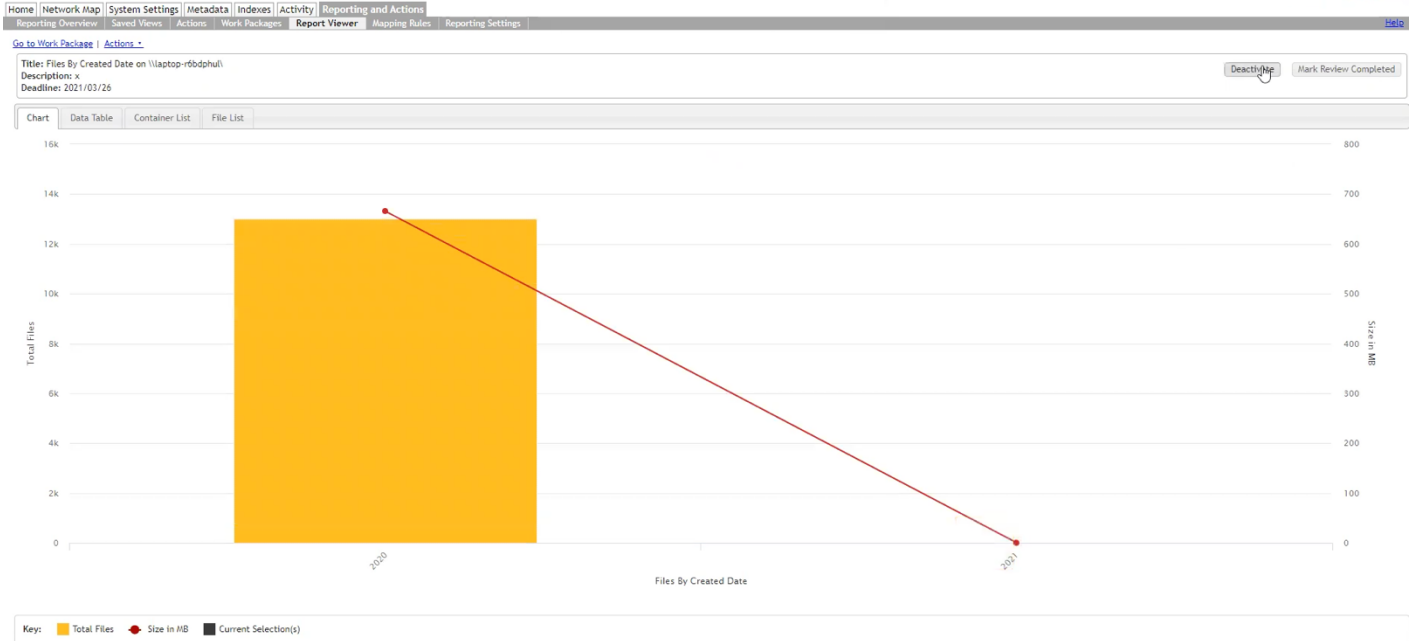


Figure 63 Deactivating a Work Package via Report Viewer

**Note.** The Report for the Work Package is displayed when the Work Package is activated; the Work Package can also be deactivated from this view using the **Deactivate** button on the top right of the screen.

1. Select the **Work Packages** tab.
2. Locate the Work Package to be deactivated and click on the launch  action icon in the Actions column of the grid.
3. The Report will now open in the Report Viewer; click the **Deactivate** button on the top right of the screen.

The Work Package is now deactivated.

## Resetting

An Information Manager can reset an **Active** or **Completed** Work Package, changing the status to **Assigned** for the Reviewer/Group.

**Note.** An Information Manager can only reset other Reviewer's Work Packages; they cannot reset Work Packages that they are the assigned Reviewer. If an Information Manager is the assigned Reviewer, they can deactivate the Work Package instead.

To reset a Work Package:

The screenshot shows the Work Packages Tab interface. At the top, there are navigation tabs: Home, Network Map, System Settings, Metadata, Indexes, Activity, Reporting and Actions. Below these are sub-tabs: Reporting Overview, Saved Views, Actions, Work Packages, Report Viewer, Mapping Rules, Reporting Settings. The main content area shows a table of work packages. The table has columns: Title, Assigned User, Active User, Owner, Status, Deadline, Completed, Total Results, Report Status, and Actions. The table contains five rows of work packages.

Title	Assigned User	Active User	Owner	Status	Deadline	Completed	Total Results	Report Status	Actions
Age By Created Date on \\laptop-r6bdphul	LAPTOP-R6BDPHUL/REVIEWERS	LAPTOP-R6BDPHUL/Reviewer1	LAPTOP-R6BDPHUL/Infomanager	Active	2021/03/26	-	13,036		
Markup on \\laptop-r6bdphul	LAPTOP-R6BDPHUL/Infomanager	LAPTOP-R6BDPHUL/Infomanager	LAPTOP-R6BDPHUL/Infomanager	Active	Due in 3 days)	-	2		
Files By Created Date on \\laptop-r6bdphul	LAPTOP-R6BDPHUL/REVIEWERS	LAPTOP-R6BDPHUL/Infomanager	LAPTOP-R6BDPHUL/Infomanager	Assigned	2021/03/26	Yesterday	-		
Files By Owner on \\laptop-r6bdphul	LAPTOP-R6BDPHUL/REVIEWERS	LAPTOP-R6BDPHUL/Infomanager	LAPTOP-R6BDPHUL/Infomanager	Assigned	2021/03/26	-	13,036		
File Extensions on \\laptop-r6bdphul	LAPTOP-R6BDPHUL/REVIEWERS	LAPTOP-R6BDPHUL/Infomanager	LAPTOP-R6BDPHUL/Infomanager	Completed	2021/03/19	Today	13,036		

Figure 64 Work Packages Tab

1. Select the **Work Packages** tab.
2. Select the Work Package from the list to be reset by clicking on its status; the status must be **Active** or **Completed**.
3. The Review Package Status screen will now appear; click on the **Reset Work Package** button.



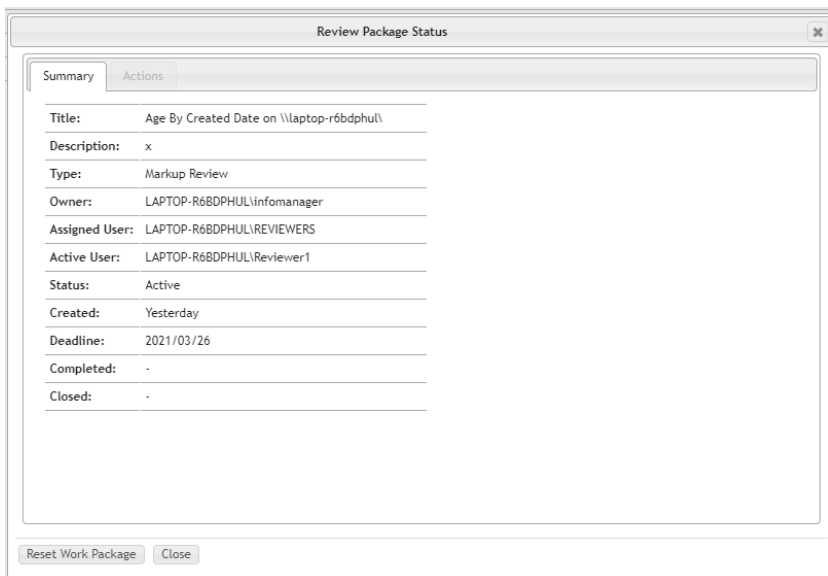


Figure 65 Resetting a Work Package

4. Confirm that you wish to proceed by clicking **OK**.

## Deleting

An Information Manager can delete a Work Package. To delete a Work Package:

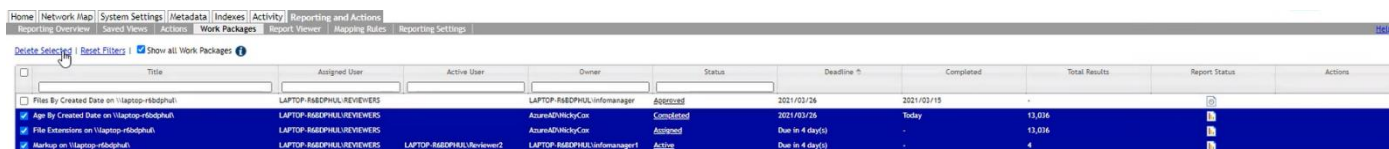


Figure 66 Deleting a Work Package

1. Select the **Work Packages** tab.
2. Check the box(es) next to the name of the Work Package(s) to be deleted or select all if you wish to delete all Work Packages.

**Note.** Approved Work Packages cannot be deleted; only Work Packages with the status Completed, Active, or Assigned can be deleted.

3. Click on the **Delete Selected** link.
4. Confirm the deletion by clicking **OK** when prompted.

**Note.** Details of any report actions taken on these packages will still be available in **Activity History**.



## Filtering

To filter the list of Work Packages:

1. Type the text string into the boxes at the top of the *Title*, *Assigned User*, *Active User*, *Owner*, and/or *Status* columns.
2. Press the Return key.
3. The Work Packages list is restricted to items that contain the specified text string(s). Click on the **Reset Filters** link at the top of the page to restore the full list.

## Sorting

Click on the column headers to reorder the list. Click on the column header a second time to reverse the sort order.

<input type="checkbox"/>	Title	Assigned User	Active User	Owner	Status	Deadline	Completed	Total Results	Report Status	Actions
<input type="checkbox"/>	Files By Created Date on \\laptop-r6bdphul\	LAPTOP-R6BDPHUL\Reviewers	LAPTOP-R6BDPHUL\Reviewer1	LAPTOP-R6BDPHUL\Infomanager1	Active	2021/03/26	-	13,036		
<input type="checkbox"/>	Files By Last Accessed Date on \\laptop-r6bdphul\	LAPTOP-R6BDPHUL\Reviewer2	LAPTOP-R6BDPHUL\Reviewer2	LAPTOP-R6BDPHUL\Infomanager1	Active	2021/03/25	-	13,036		
<input type="checkbox"/>	Files By Owner on \\laptop-r6bdphul\	LAPTOP-R6BDPHUL\Reviewer1	LAPTOP-R6BDPHUL\Reviewer1	LAPTOP-R6BDPHUL\Infomanager1	Assisted	Tomorrow	-	13,036		
<input type="checkbox"/>	File Extensions on \\laptop-r6bdphul\	LAPTOP-R6BDPHUL\Reviewer2	LAPTOP-R6BDPHUL\Reviewer2	LAPTOP-R6BDPHUL\Infomanager1	Assisted	2021/03/25	-	13,036		

Figure 67 Work Packages tab

## Report Viewer

When you first access the *Report Viewer* tab, Discovery Center will prompt you to define a new view. You can also start a new report by clicking on the **Define view** link on the *Saved Views* tab.

The *Report Viewer* tab is displayed by Discovery Center whenever you:

- Create a new view in the *Define View* dialog box (see below).
- Select an existing chart from the *Saved Charts* dropdown list on the *Areas of Interest* page (see page 109).
- Select an existing view from the *Saved Views* tab (see page 114).

If you manually select the *Report Viewer* tab, you are prompted to create a new view in the *Define View* dialog box (see **Creating a Report**, below).

## Creating a Report

**Note.** Ensure that the reporting database has been recently processed by your AN Administrator. If the reporting database has not been processed, a saved view may not truly reflect the state of current data.

You can start a new report by clicking on the **Define view** link on the *Saved Views* tab, directly on the *Report Viewer* tab or as an *Action* from the *Reporting Overview* tab. If you are viewing an existing report, click on the **Define View** button. Choose from the following options:

### 1. Report Type

Choose the type of report (see page 128):

- **Calculated Fields**  
If you choose this option, select the Calculated Field that you want to use from the displayed dropdown list.
- **Containers**  
Displays a ring chart depicting the folder structure in the chosen Area of Interest or Network Location. Optionally, add a Field heat map overlay:



- **Show Field Heat**  
Select a field (basic metadata or calculated field) to overlay the chart with a heat map showing its distribution and concentration within containers. Choose how to normalize the 10-color heat map:
- **Matched File Count**  
Normalize the heat map according to the number of files that exhibit one or more values for the selected field.
- **Intensity**  
Normalize the heat map according to the number of hits recorded in a container (including multiple values in the same file) for the selected field.
- **Content and File Duplicates**  
**File Duplicates**  
If you choose either of these options (see **Duplicates Reports**, page 133, for more information), select how Discovery Center identifies master documents using the *Master Selection Strategy* dropdown list:
- **Earliest Created**  
In this type of report, the file in each duplication cluster with the earliest date created property is identified as the *master* file and all others are labeled as *File Duplicates*.
- **Last Modified**  
In this type of report, the file in each duplication cluster with the most recently saved changes is identified as the *master* file and all others are labeled as *File Duplicates*.
- **Area of Interest**  
Using the **Master Area Of Interest** dropdown list, identify the AOI where master documents are stored. The selected location for the report must not be within, or the same as, the Master Area of Interest.

**Note.** If the selected location for the report does not include the Master AOI there will not be a Master segment on the pie chart: in this case, the master documents are, by definition, outside of the reporting scope.

- **File Extensions**  
Displays a vertical bar chart of files by extension. Note that large servers may contain many thousands of file extensions, making this chart difficult to read; filter the chart to reduce the number of extensions returned.
- **Files by Created Date**  
Displays a vertical bar chart of files by Date Created metadata.
- **Files by Last Accessed Date**  
Displays a vertical bar chart of files by Date Last Accessed metadata.  
**Note.** The accuracy of *Date Last Accessed* metadata depends on your local server configurations and policies. It is often not a reliable indicator of the last access time of a file by a user: Versions of Windows currently supported by Microsoft do not update this attribute at all by default; viewing the file properties using Windows Explorer can cause an update to this attribute if it is enabled.
- **Files by Modified Date**  
Displays a vertical bar chart of files by Date Last Modified metadata.
- **File Owners**  
Displays a vertical bar chart of files by owner. Note that large servers may contain many thousands of file owners, making this chart difficult to read. Filter the chart to reduce the number of owners returned.

## 2. Selected Location

Identify the Area of Interest or Network Location you want to investigate.

## 3. Filters

Optionally, apply filters to the chart data. You can filter by:

- **File Age**  
Filter the data according to the Created Date, Modified Date or Last Accessed Date.
- **File Size**  
Set a range of file sizes (B, KB, MB or GB) to include in the report.



- **Calculated Fields**

Filter the data according to the discovered metadata. Enable the **Filter by calculated field** check box and choose the *Calculated Field and Value (Has a value, No value or Custom values)*.

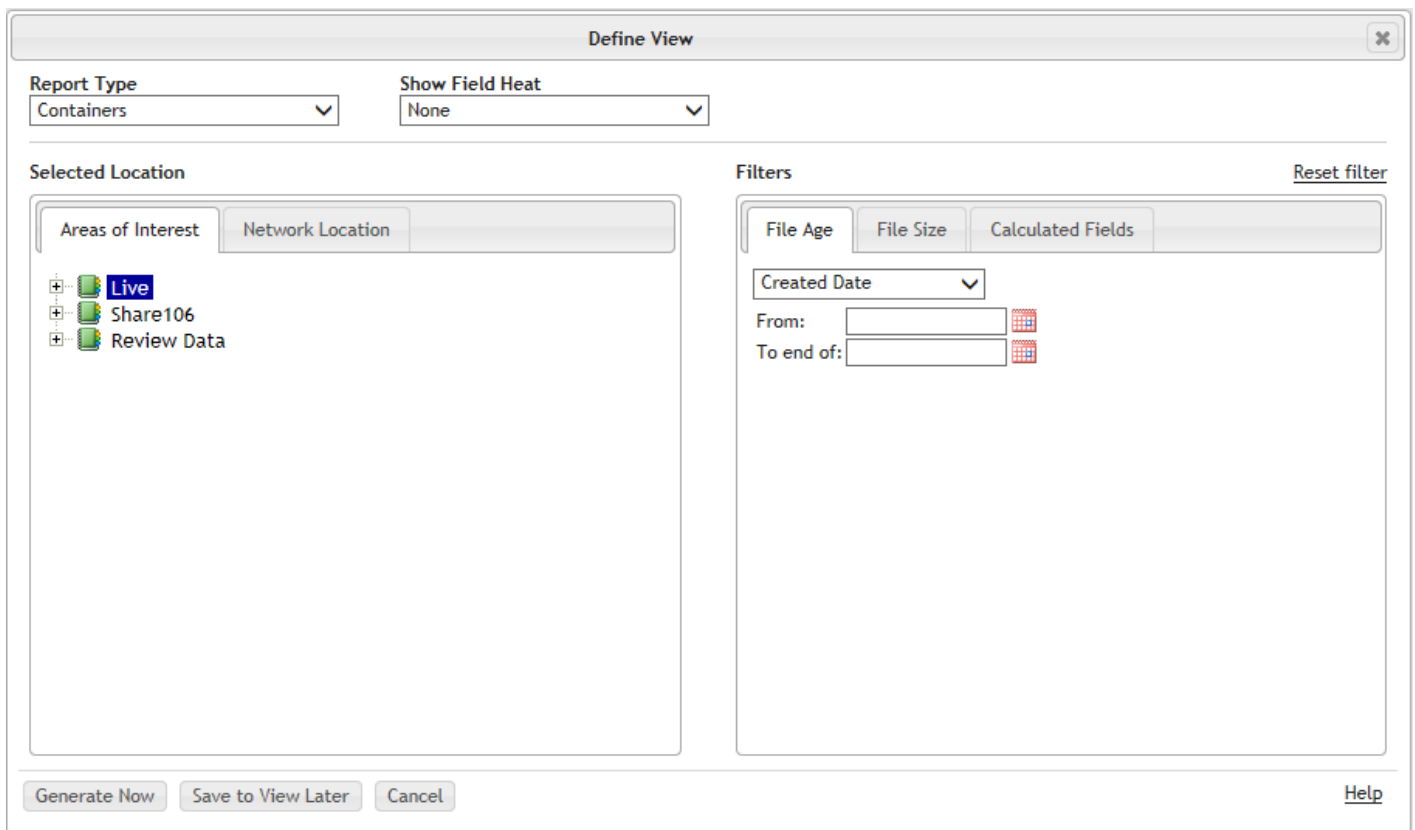


Figure 68 Define View dialog box

Click on the **Reset filters** link if you want to remove all enabled filter options.

4. Click on **Generate Now** to create and display the view.

Alternatively, if you are defining several views and want to avoid waiting for each view to be generated, you can save time by clicking on **Save to View Later**. Enter a name for the view when prompted. The view will be listed in the *Saved Views* tab (see page 114).

**Note.** Reporting performance depends upon SQL Server hardware specification, the number of files in the report and any other tasks being carried out by the Discovery Center host. Reports across millions of files may take several minutes to produce. The addition of multiple filters will also reduce reporting performance.

When the report has been compiled, the chart and associated summary data is displayed on the *Reporting and Actions* page's *Report Viewer* (see page 125).



# Types of Report

## Calculated Fields Report

A Calculated Fields report displays a bar chart showing the values of metadata matching the chosen Calculated Field. Files without any corresponding metadata are collated in a No value category.

This type of report allows you to explore calculated fields to profile and act on indexed information. In addition to any custom calculated fields you may have created, certain types of calculated field provide important reporting functionality; these fixed fields are added to all indexes and cannot be removed. Where necessary, their definitions can be altered by editing and uploading the supporting classification file.

The following fixed fields are available:

- **Age By Created Date**  
Classifies files according to their age as calculated from the *Created Date*. The value is applied at the time that the document is classified, and so represents the age of the document at the time of classification based on the last modified date read for the file at the time the index skim was last run.
- **Age By Last Accessed Date**  
Classifies files according to their age as calculated from the *Accessed Date*. The value is applied at the time that the document is classified, and so represents the age of the document at the time of classification based on the last accessed date read during the preceding skim.  
**Note.** On a Windows file system, the Last Accessed Date property might be updated by background tasks (such as anti-virus scans) or not updated at all depending on the server configuration. It may not therefore accurately represent the last time a document was accessed by a user for viewing, printing or other purposes.
- **Age By Modified Date**  
Classifies files according to their age as calculated from the *Modified Date*. The value is applied at the time that the document is classified, and so represents the age of the document at the time of classification based on the last modified date read for the file at the time the index skim was last run.
- **File Sizes**  
Classifies files in labelled size ranges.
- **Filetypes By Extension**  
Classifies files by purpose (spreadsheets, image files, word processing, etc.) according to their file extension. The classification rules for this field include more than 4000 file extensions. Note that approximately 25% of the extensions classify into more than one file type.  
**Note.** In applications where you need up-to-date information, particularly from reports using Age By... fields, ensure that the relevant indexes are run on a regular schedule (see page 174 for details).





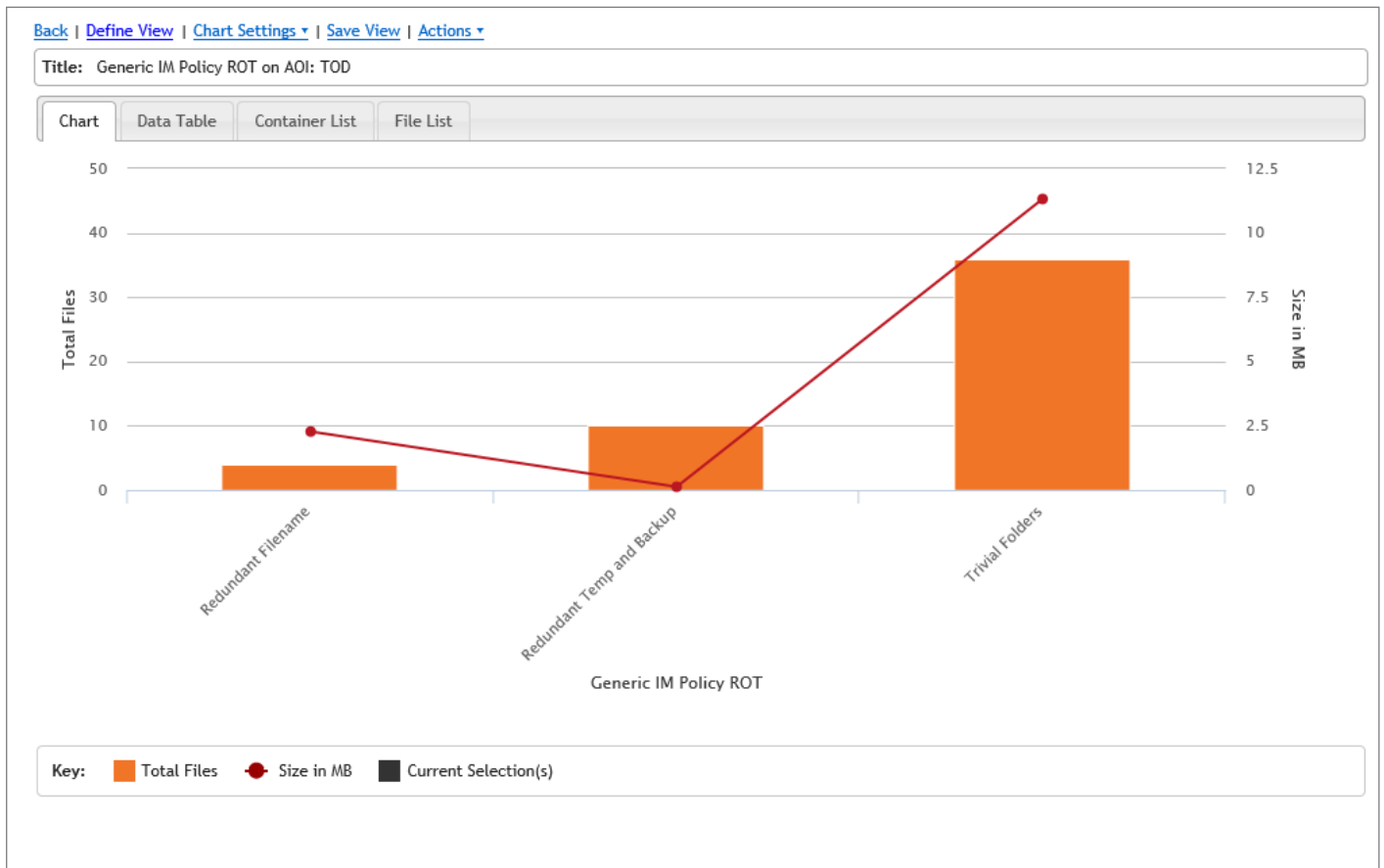


Figure 69 Calculated Fields Report

**Note.** Charts are limited to 30 columns. If your data contains more than 30 different categories, the smallest data sets are collated and shown in a column labeled "Other". This column cannot be selected; use the *Data Table* tab to investigate the data sets that contribute to this column.



## Containers Report

Use a Containers Report to explore the distribution of information in your file store and identify where the folder structure is poorly configured, where folders are not being used or are too heavily populated. By applying a 'field heat' overlay, you can highlight the locations of target files by exploring the relative concentrations of discovered metadata or calculated field values in different containers.

A Containers Report consists of a ring or multilevel pie chart in which hierarchical data is depicted by concentric circles. The circle in the center represents the Area of Interest or Network Location chosen for the report. Segments in the next ring represent the Containers (folders or zip files) within the Area of Interest/Network Location and their relative contribution in terms of either size, or number of files (as selected in *Chart Settings*). Moving out from the center of the chart, segments which lie within the angular sweep of an inner segment bear a corresponding hierarchical relationship to it. The size of a segment includes the contribution of content that is held immediately in it and the contribution of all of the child containers. Content contained directly in a parent segment results in corresponding gaps between the child segments.

To find out more information about a segment, hover the mouse pointer over it. A popup box shows the container's file path, the number of files and child containers it contains, and the total size of all contained files.

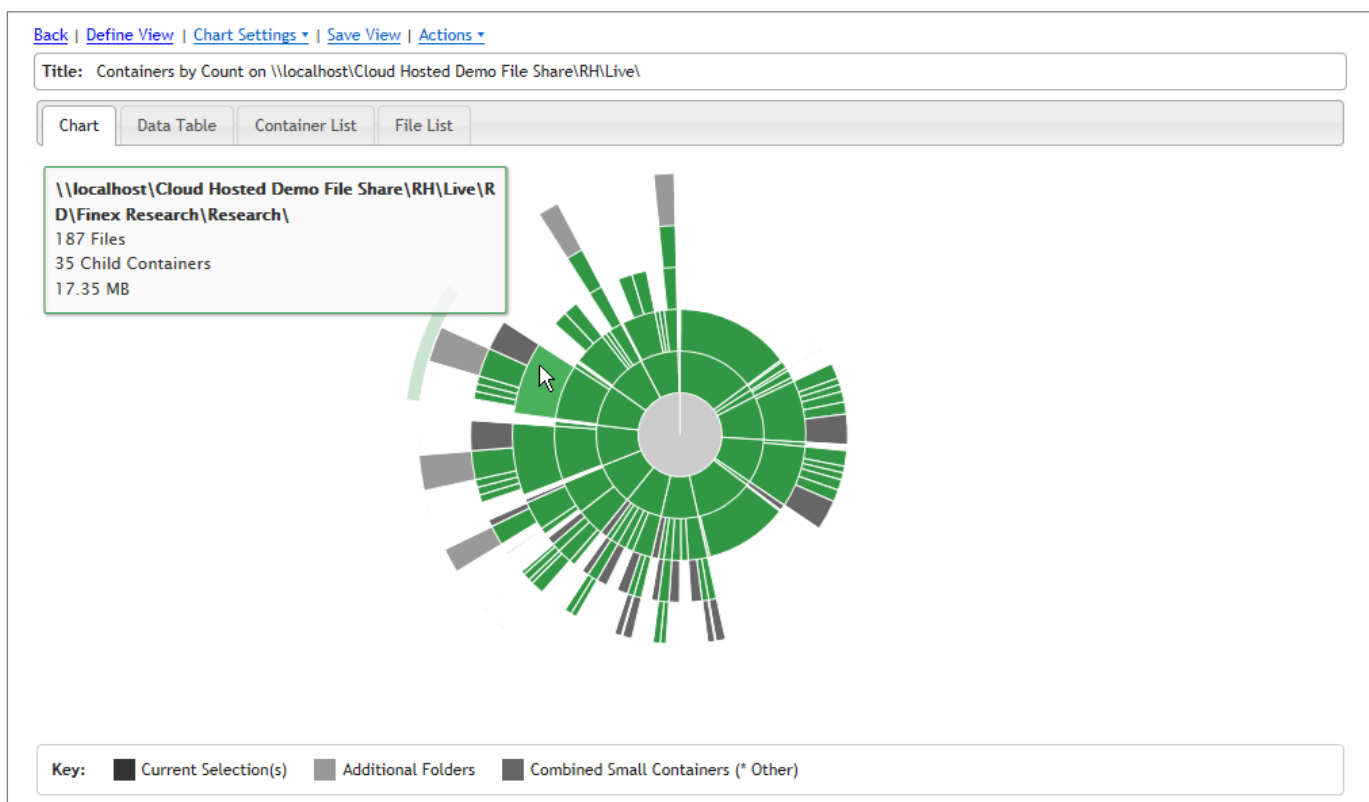


Figure 70 Containers Report (without a Field Heat overlay)

### Drilldown

An additional feature of the Containers Report is the ability to drill down to examine a selected 'segment'. Each chart is only capable of showing a limited number of hierarchical layers, so the drilldown feature allows you to examine deeply nested structures. Select the container you want to investigate, then right-click on the chart and select **Set view location** from the context menu. Alternatively, select the command from the *Actions* menu.



## About the Key

Container segments are colored green unless you choose to apply one or more File Age, File Size or Calculated File filters when defining the report (see

Creating a Report, page 125 and *Using Field Heat*, below). Segments are colored gray in the following circumstances:

- **Additional Folders**  
A Containers Report can show multiple hierarchical layers of containers depending on the value of the Container Chart Depth setting. If the structure of the location is more extensive, segments in the outer ring are given a light gray color. To explore these containers, select the parent container and use the Set view location command (see *Drilldown*, page 130).
- **Combined Small Containers (\* Other)**  
A darker gray segment indicates that 2 or more values have been aggregated into a single section of a chart. This happens in charts with segments that are less than 1/200th of the total data presented.
- **Current Selection(s)**  
When you click on one or more segments, they become enabled for commands on the Action menu. To indicate this, the segments are given a dark gray color. Click on another segment or elsewhere on the chart to deselect the original selection and restore the normal coloration.

## Using Field Heat

When you define a Containers Report, you are given the option to apply one or more File Age, File Size or Calculated File filters (see

Creating a Report, page 125). This allows you to investigate the distribution of calculated fields within containers. If one or more filters have been applied, these are listed in the chart header, under the title. Containers are colored from blue to red according to either the number or percentage of matched files. Blue identifies containers with no matches for the selected filters. By default, the colored scale of the overlay is based on the number of matched files normalized to the highest number of matches in any of the displayed containers. The overlay can also be based on percentage: the proportion of files in each container matching the filter criteria. To use this display option, select *Chart Settings* and choose *Show Container Heat By Percentage*.

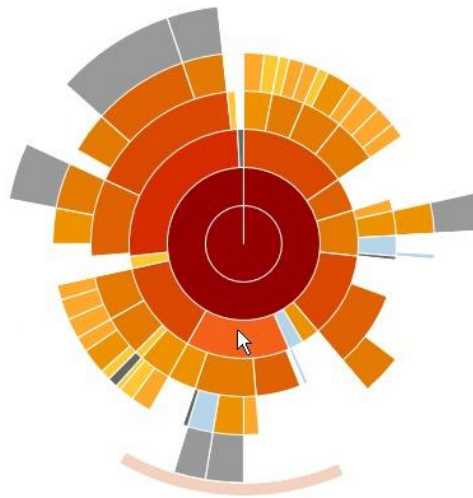


[Back](#) | [Define View](#) | [Chart Settings](#) | [Save View](#) | [Actions](#)

Title: Containers by Count on AOI: TOD  
Filter: Age By Modified Date has custom values.

Chart Data Table Container List File List

\\localhost\test old docs\temp\  
34 Files  
26 Child Containers  
14.79 MB  
Matched File Count: 26 (76%)



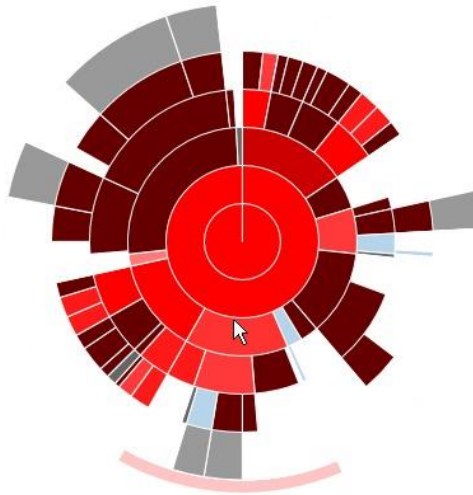
Key: ■ Current Selection(s) ■ Additional Folders ■ Combined Small Containers (\* Other) ■ Matched File Count (0 to 203)

[Back](#) | [Define View](#) | [Chart Settings](#) | [Save View](#) | [Actions](#)

Title: Containers by Count on AOI: TOD  
Filter: Age By Modified Date has custom values.

Chart Data Table Container List File List

\\localhost\test old docs\temp\  
34 Files  
26 Child Containers  
14.79 MB  
Percentage Matched Files: 76% of 26 files



Key: ■ Current Selection(s) ■ Additional Folders ■ Combined Small Containers (\* Other) ■ Percentage Matched Files (0 to 100)

Figure 71 Containers Report with Field Heat overlay

In this case, the overlay represents hits for files matching a defined *Modified Date* range based on file count (above) and percentage (below).



## Duplicates Reports

### About File and Content Duplicates

Microsoft Office documents (DOC, DOCX, PPT, PPTX, etc.) consist of two main components: a metadata header and the file content. When a file is copied from one folder to another, the process creates an exact copy of the file in terms of both content and metadata. In Discovery Center, if one of these files is identified as the master, the other file is termed a **File Duplicate**. You can use a *File Duplicates Report* to assess this form of duplication in an Area of Interest or network location. However, if an Office document is exported to SharePoint, some of its metadata fields may be altered during the process. For example, the *Created* and *Modified* dates are set automatically by SharePoint to the date and time that the file was moved. The file may also acquire other fields from the destination document library.

The SharePoint version of the file will no longer be the same as the original even though their file content is the same. Each time a copy of an Office file passes into SharePoint its metadata may diverge from that of the original file and other versions stored elsewhere. In Discovery Center, a file with the same content as the corresponding master but with different metadata is called a Content Duplicate. A *Content and File Duplicates* report detects this form of duplication. In a File Duplicates report, Content Duplicates would not be identified as copies of the master file.

**Notes.** Other file types, such as PDFs or image files, are not modified by SharePoint in this way. However, Content Duplicates of any file may arise from the manual editing of file metadata. Files transferred to and from other file repositories may be subject to metadata changes.

### Cleansing Duplication

You can use *File Duplicates* and *Content and File Duplicates* reports to cleanse file stores. The reports identify File Duplicates (exact matches) and Content Duplicates (files with identical content but differing metadata) but cannot highlight files with similar content.

If your network map includes SharePoint locations, or Office files that have been previously stored in SharePoint, you should use the *Content and File Duplicates* Report. In a network without SharePoint locations, where there is unlikely to be a significant number of Content Duplicates, use the *File Duplicates* report. This is less intensive in terms of processing time. You could run a *Content and File Duplicates* report on a small sample of the network location to confirm that this is the case.

### Using the AOI Master Selection Strategy

When you create a Duplicates Report, you must specify a *Master selection strategy*. There are three options:

- First Created
- Last Accessed
- Area of Interest.

When you select the *Area of Interest* option, you will be prompted to identify an Area of Interest that defines the location where master documents are stored. The selected AOI is then used as a reference and is compared with the content in AOI or network location selected for the report. If the location for the report does not include the location(s) identified by the Master AOI then there will never be a Master segment on the pie chart; in this case the master documents are, by definition, outside of the reporting scope. If you see a Master segment on the chart for this strategy then this indicates that your reporting location includes some or all of the locations defined in the Master AOI. This is illustrated in the examples below.



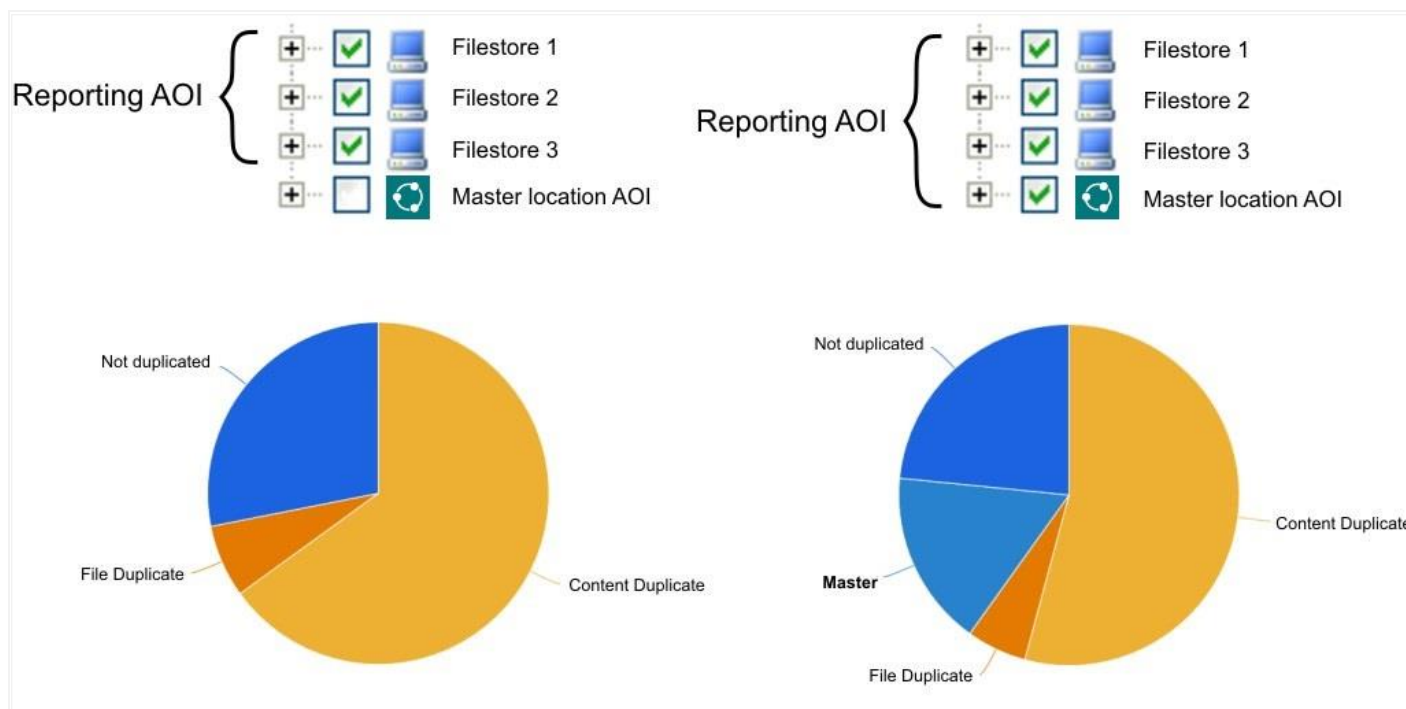


Figure 72 Using the Area of Interest master selection strategy

Example of a Content and File Duplicates report run with the master location outside (left) and inside (right) the scope of the report

You cannot create a duplication report for an AOI or Network location that is totally within the chosen Master AOI; no duplication would be identified. Documents in an AOI used as a Master location will never be marked as duplicates because they are by definition, 'Master' documents, even if there are multiple copies in the Master AOI.

If there are duplicate documents in the Master location and the Master location partially overlaps the report location, then it is possible for groups of duplicates to be identified that have more than one master document. If duplicates from outside of the Master AOI are removed when there are multiple possible Master documents then it is indeterminate which document will be linked with a shortcut, so it is preferable to de-duplicate any AOI used as a Master location.

#### Dealing with Duplication

The following procedure highlights the general workflow for cleansing a filestore of duplicate files using reports and reporting actions. Note that the removal of redundant or sensitive files is a separate process. Beforehand, and after delete or migrate actions, all network locations must be indexed and analyzed, and the reporting database updated.

1. **Create or identify a location where you want to store master copies.**  
This may be another network location or a repository such as SharePoint.
2. If there are files already present in the master location, check it for duplication.  
**Create an Area of Interest for the master location and define a Duplicates Report using a *First Created or Last Accessed* strategy. If any duplicates are present, use the Delete action to quarantine the copies or the Migrate option to archive the files elsewhere.**
3. Search for duplicates of the master files across the remaining network locations.  
**Create one or more AOIs to cover the network locations of interest. Define a Duplicates report for each of these AOIs using the master location's AOI as the Master selection strategy (see below). If any duplicates are present, use the Delete or Migrate actions to remove them.**

4. Search for files to migrate to the master location.

**Using the AOIs for the locations to be cleansed, define a File Duplicates Report using a First Created or Last Accessed master selection strategy. Use the Delete action to quarantine duplicates. Regenerate the index and report. Use the Migrate action to transfer or copy the remaining unique files to the master location.**

#### File Duplicates Report

A File Duplicates report allows you to examine the extent of file duplication in an Area of Interest or a network location. File Duplicates are exact binary copies, with matching content and internal metadata. If two files have differing file system metadata (for example, different Last Accessed dates) they will be considered identical, but Microsoft Office file formats that store metadata internally will not be identified as duplicates in this report.

The report is a pie chart with one or more of the following segments:

- **Master**  
Files identified by your chosen master selection strategy: First Created, Last Modified or by location (Area of Interest).
- **File Duplicates**  
Exact binary matches of the corresponding Master files.
- **Not duplicated**  
Files without any copies in the context of the current report.
- **Empty**  
Zero byte files. Often these files can be deleted but sometimes information might be contained in the filenames.
- **Unanalyzed**  
Files that have not had Duplicate Analysis applied.

As with other reports, the *Files List* includes files from each selected segment of the report, and all files from the report location if there are no selections made. If you select both the *File Duplicates* and *Master* segments, the Files List shows sample clusters of duplicate files.



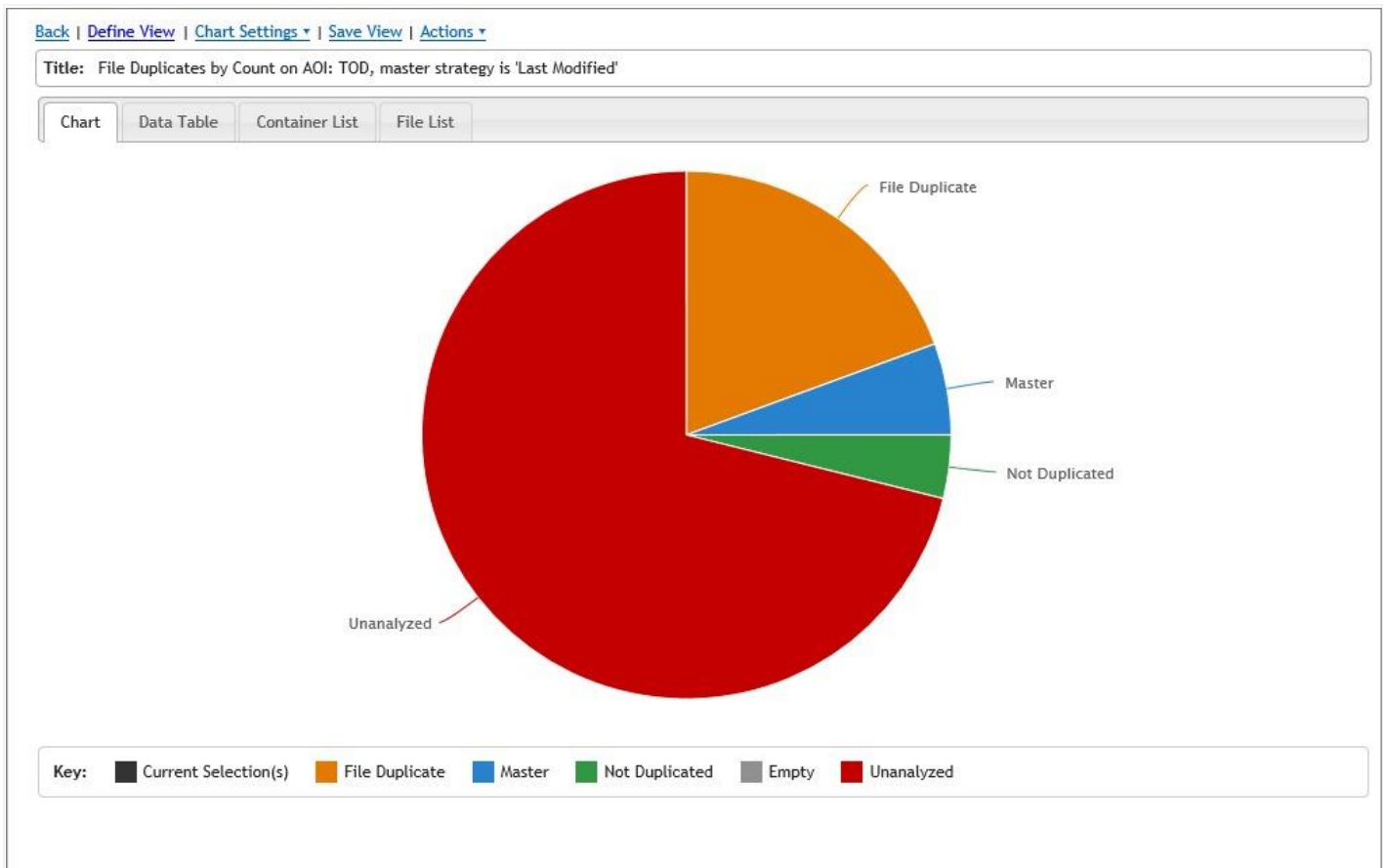


Figure 73 Example of a File Duplicates report

#### Content and File Duplicates Report

A Content and Files Duplicates report allows you to examine the extent of duplication in an Area of Interest or a network location containing one or more SharePoint locations. This report will identify files that are exact 'binary' duplicates (File Duplicates) and those that share identical content but differing metadata (Content Duplicates).

The report is a pie chart with one or more of the following segments:

- **Master**  
Files identified by your chosen master selection strategy: First Created, Last Modified or by location (Area of Interest).
- **Content Duplicates**  
Files whose content is identical to that of the corresponding master files but with one or more differing metadata fields.
- **File Duplicates**  
Exact binary matches of the corresponding Master files (identical content and metadata).
- **Not duplicated**  
Files without any copies in the context of the current report.
- **Empty**  
Zero byte files. Often these files can be deleted but sometimes information might be contained in the filenames.
- **Unanalyzed**  
Files that have not had Duplicate Analysis applied.





As with other reports, the *Files List* includes files from each selected segment of the report, and all files from the report location if there are no selections made. To populate the *Files List* only with groups of duplicated files, select the *Master*, *File Duplicates*, and *Content Duplicates* segments.

**Note.** Content duplicate analysis is enabled by default but can be disabled for an index (see page 92).

If content duplication is disabled for some folders in a report (either a folder in the report location or in an AOI used in the Master Selection Strategy) and the documents in that folder are part of any group of file duplicates then the system will detect this partial content analysis and will not generate the report. In this case, either re-analyze the folder with content duplication enabled or remove the partially analyzed locations from the report.

**Notes.** The *Content and File Duplicates* report is most often useful when one or more of the selected Areas of Interest includes a SharePoint location. If this is not the case, then the *File Duplicates* report is likely to identify the same amount of duplication with faster report generation. Changing between date-based Master Selection Strategies can have a significant effect on the way that duplication is identified by the *Content and File Duplicates* report, depending on the relationship between the master file and the remaining duplicates.

For the date-based Master Selection Strategy, files are classified according to the following logic:

- Identify all groups of documents that match exactly or by document content
- Determine which document in each group of duplicates is the master according to the selected date based strategy
- Files which are File Duplicates of a master document are reported as File Duplicates
- All other duplicated files are reported as Content Duplicates although there may be File Duplicates within this group

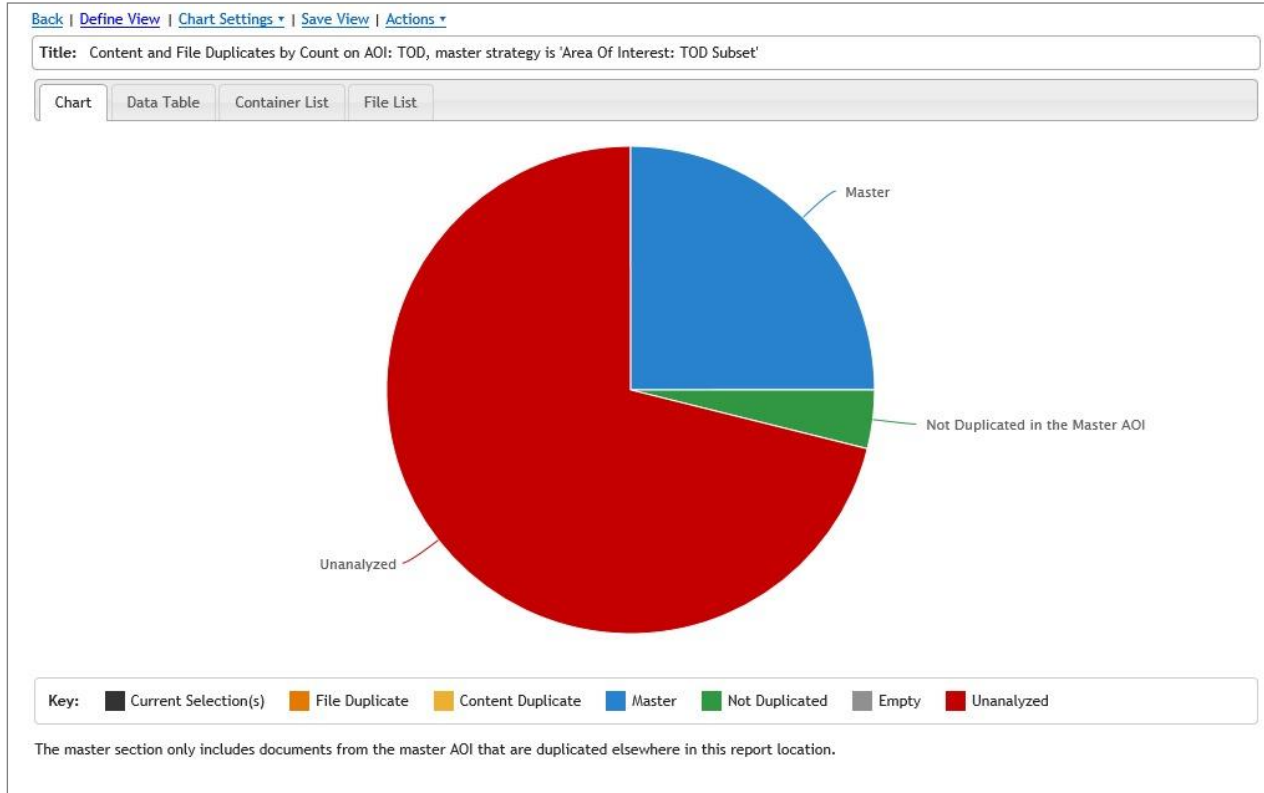


Figure 74 Example of a Content and File Duplicates report



## File Extensions Report

This report type describes the frequency distribution of files in the AOI or location according to file extension. The data is shown as a bar (number of files) and line (file size) chart.

Any files without an extension are collated in a No extension data set.

Use a File Extensions Report in the early stages of a project to assess the nature of files stored in any environment. File extensions are well-understood and so this report is easy to present to those new to information projects and activities. Note that there are many thousands of known file extensions and so large file shares can produce very complex reports. Apply filters to focus on specific areas. For example, the *file type by extension* filter can help focus on specific file types such as spreadsheets or images, or you could filter by *no value* to show unclassified files or those where the extension has been corrupted.

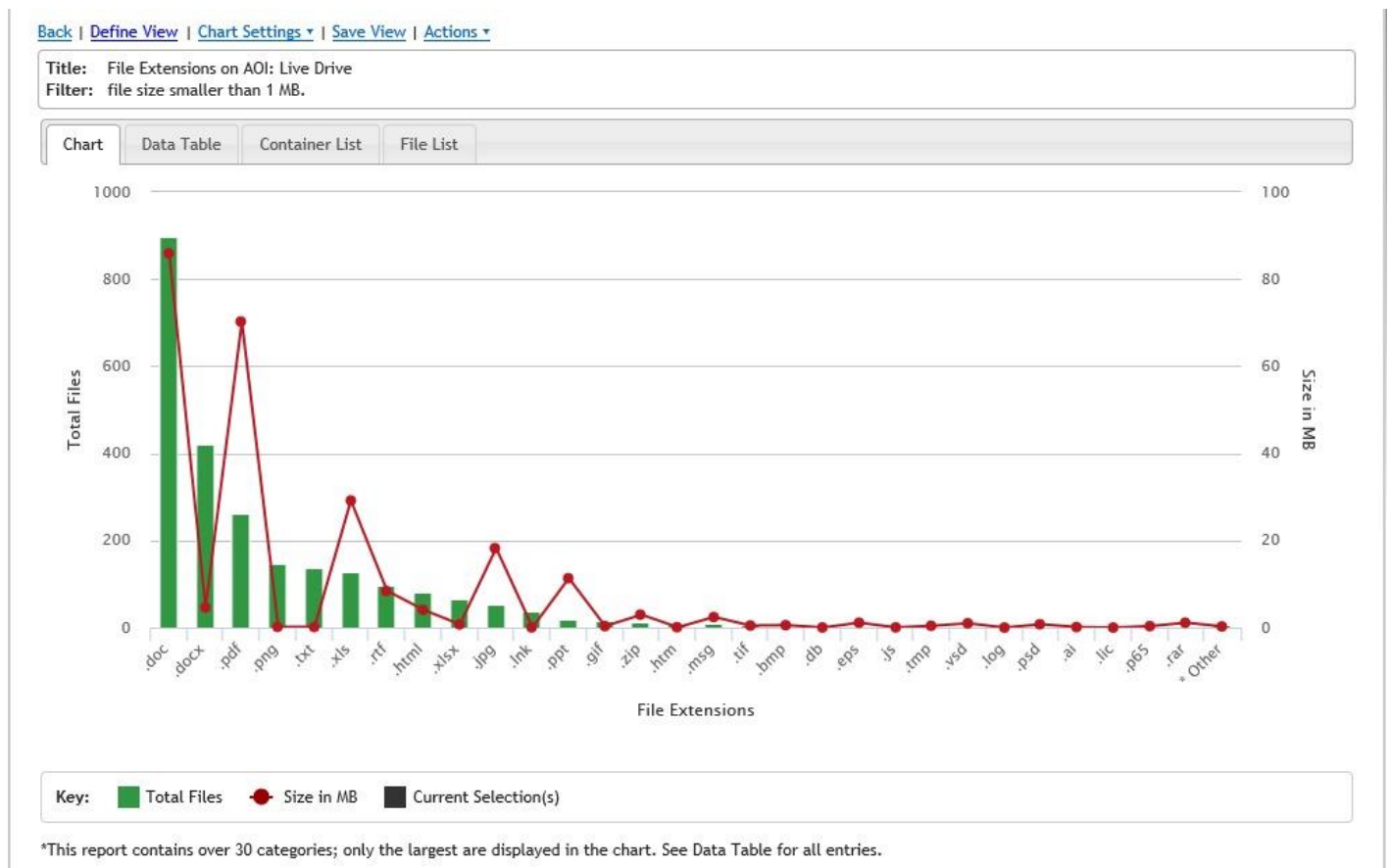


Figure 75 File Extensions Report

**Note.** Charts are limited to 30 columns. If your data contains more than 30 different categories, the smallest data sets are collated and shown in a column labeled "Other". This column cannot be selected; use the *Data Table* tab to investigate the data sets that contribute to this column.



## Files By Created/Last Accessed/Last Modified Reports

These report types describe the frequency distribution of files in the AOI or location according to the year each file was:

- Created
- Last Accessed, or
- Last Modified

Any files without the corresponding metadata can be collated in a *No value* category (this is disabled by default but can be set in Chart Settings (see page 144)).



Date reports show the historical activity across indexed files. Whilst all such reports are subject to weaknesses in the way file shares capture and store dates, the patterns reveal when large changes occurred (for example, a bulk migration of files) or where information is unchanged and possibly stale. Exercise caution when using 'last accessed' dates: these dates can be set by applications, such as backup programs or virus scanners, as well as people.

**Note.** The accuracy of Date Last Accessed metadata depends on your local server configurations and policies. It is often not a reliable indicator of the last access time of a file by a user: Versions of Windows currently supported by Microsoft do not update this attribute by default; viewing the file properties using Windows Explorer can cause an update to this attribute if it is enabled.

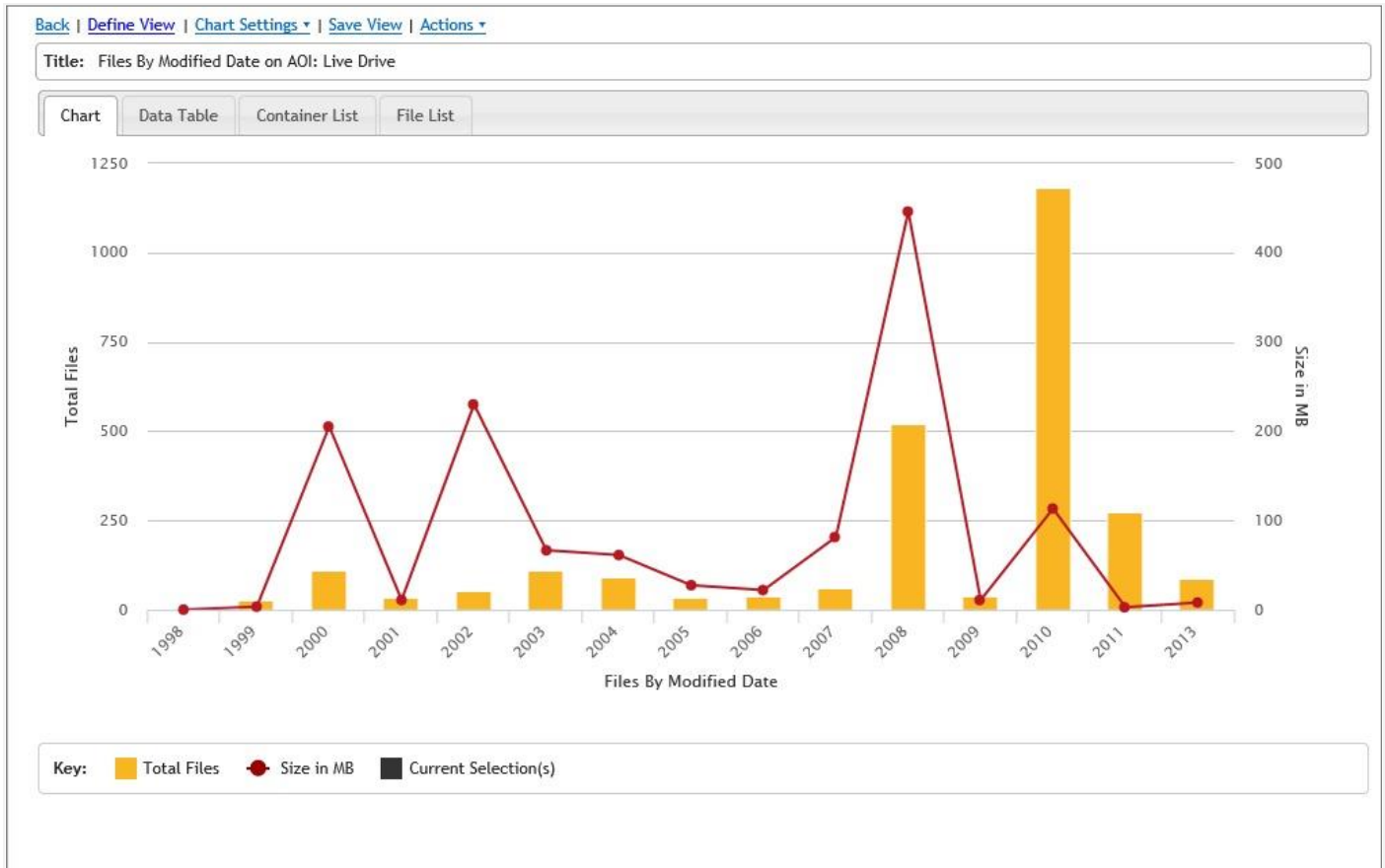


Figure 76 Files By Created/Last Accessed/Last Modified Report

**Note.** Charts are limited to 30 columns. If your data contains more than 30 different categories, the smallest data sets are collated and shown in a column labeled "Other". This column cannot be selected; use the *Data Table* tab to investigate the data sets that contribute to this column.

### Files By Owner Report

This report type displays a vertical bar chart of files by owner. Note that large servers may contain many thousands of file owners, making this chart difficult to read. Filter the chart to reduce the number of owners returned.



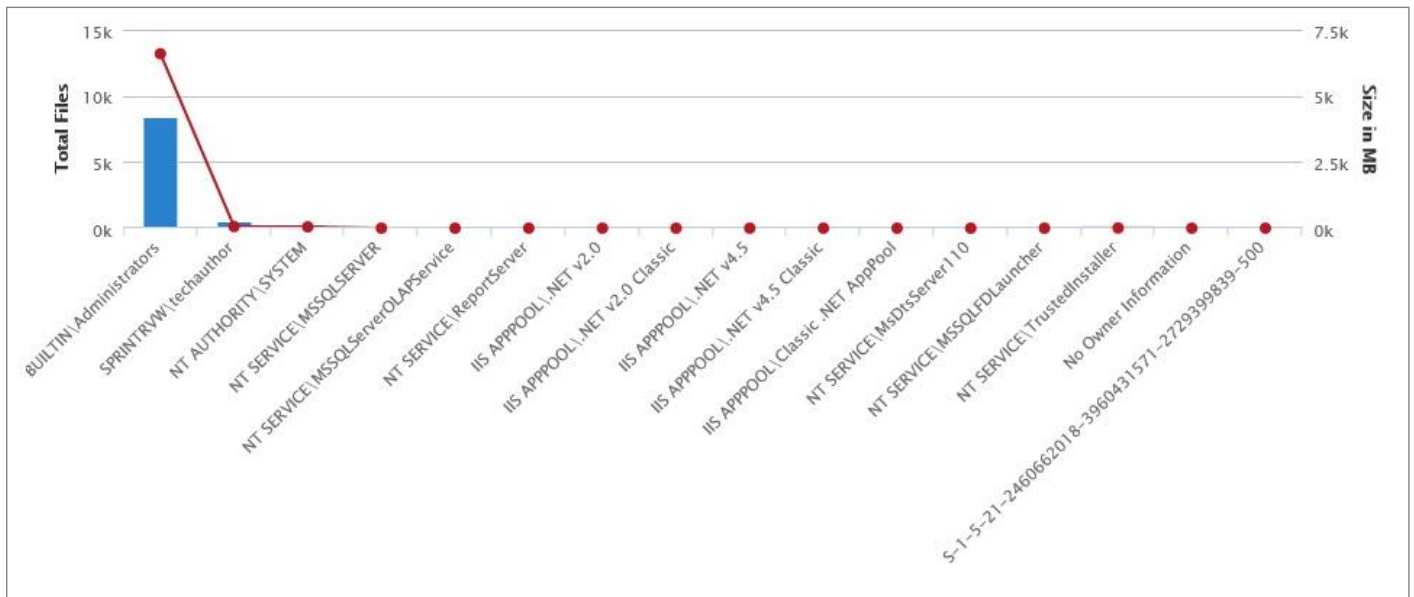


Figure 77 Files By Owner Report

**Note.** Charts are limited to 30 columns. If your data contains more than 30 different categories, the smallest data sets are collated and shown in a column labeled "Other". This column cannot be selected; use the *Data Table* tab to investigate the data sets that contribute to this column.



## Working with Reports

The *Report Viewer* page displays a summary report on the first two tabs: *Chart* and *Data Table*. These summarize index results within a location or Area of Interest and do not show file-level information. Consequently, security settings for indexes do not restrict the viewing of *Chart* or *Data Table* information. The *Container List* tab shows file locations and is subject to any filters applied when defining the report or selections on the *Chart* or *Data Table* tabs.

The *File List* tab provides access to the detailed results of indexing. Index security credentials determine, based upon Windows Authentication, if you are able to inspect files and file metadata on the *File List* tab.

To perform reporting actions (see page 151), you can select columns or pie-segments on the *Chart* tab, the corresponding data rows on the *Data Table* tab or individual files on the *File List* tab. The selections in each tab interact, so that selecting a section of a chart (column or segment of pie) will cause the corresponding row or rows to be selected in the *Data Table*, and vice versa. Unless you make selections on the *Chart* or *Data Table* tabs, the *File List* will show all files in the report location. Otherwise, the *File List* is restricted to the files in the selected *Chart/Data Table* categories.

**Note.** It is not possible to select the \* *Other* aggregation if it appears in a *Chart*, but it is possible to select the items that are rolled-up within it from the *Data Table* tab. Any selections from the *Data Table* tab will be lost if you return to the *Chart* tab.

The following function buttons are available:

- **Back**  
For the length of the browser session, Discovery Center stores the ten most recent view definitions. Press the Back button to step through the View history and regenerate a previously viewed report.
- **Define View**  
Create a new report view (see **Creating a Report**, page 125)
- **Chart Settings**  
Customize the chart (see **Customizing the Chart**, page 144).
- **Save View**  
Save the current report. You will be prompted to provide a name for the report: it will be added to the list on the *Saved Views* tab.
- **Go to Work Package** (*Displayed for reports opened from a Work Package only*)  
Display Work Package status (see page 118).
- **Actions**  
Migrate or delete files, update metadata, or export charts or tabulated data in CSV format (see **Actions**, page 151). With a *Containers* report only, there are also commands to change the view location and examine field values (see page 130). When a report is opened in Review mode from the *Work Packages* tab, *Markup Selected Fields* is the only supported action.



## Chart tab

When you first access the Report Viewer tab, Discovery Center will prompt you to define a new view (see

Creating a Report, page 125). You can also start a new report by clicking on the **Define a new report view** link on the *Saved Views* tab.

The *Chart* tab summarizes index results within a location or Area of Interest according to the chosen options in the *Define View* dialog box. There are several chart types (see page 128).

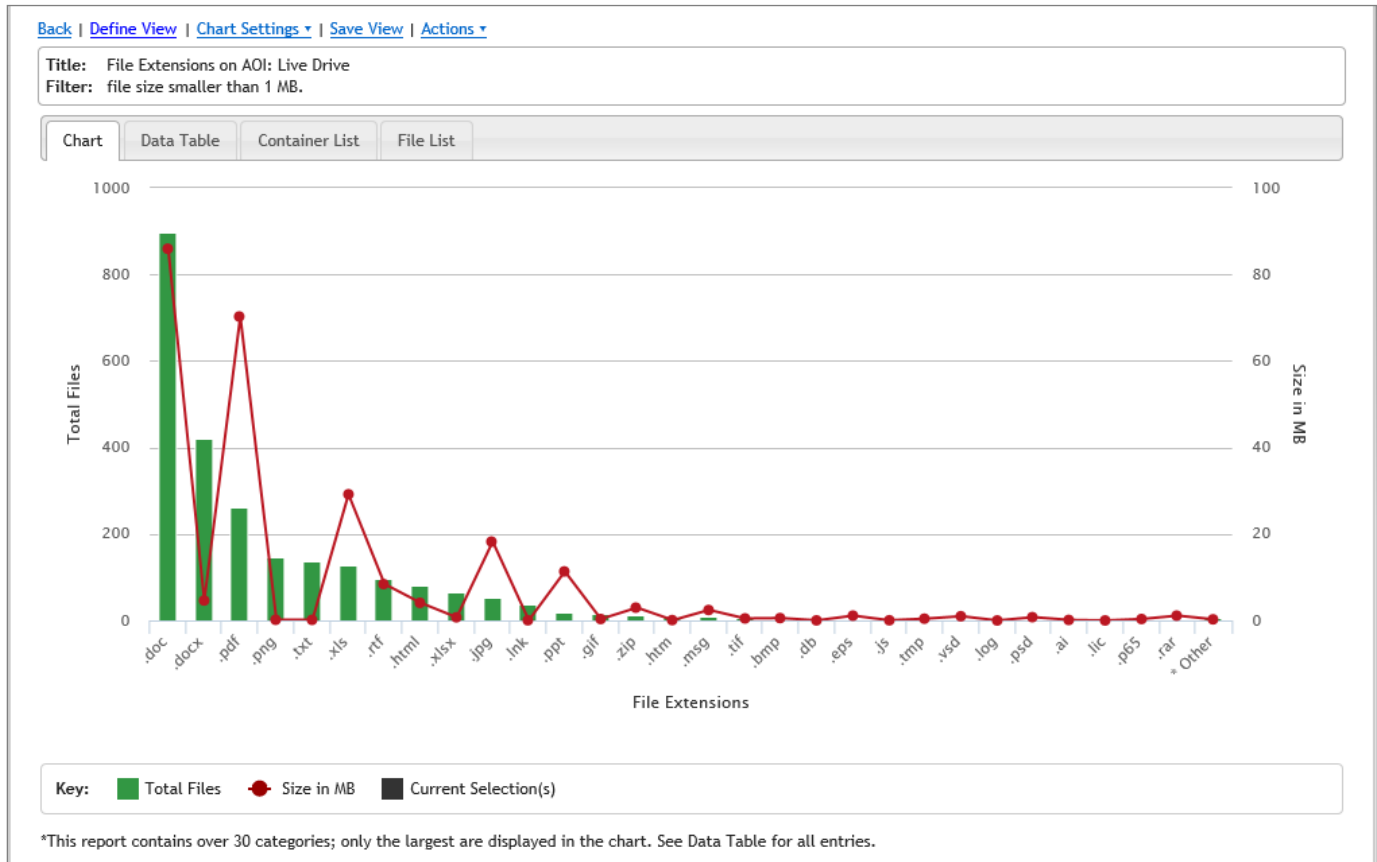


Figure 78 The Report Viewer/Chart tab

**Note.** Charts are limited to 30 columns. If your data contains more than 30 different categories, the smallest data sets are collated and shown in a column labeled "Other". This column cannot be selected; use the *Data Table* tab to investigate the data sets that contribute to this column.



## Customizing the Chart

The chart presentations can show file counts or the total size in bytes of each file included in a chart. The column charts show both quantities together, with a column chart measured against the left-hand primary axis and a line chart overlaid and plotted against a secondary axis on the right-hand side.

Click on the **Chart Settings** button to change the way the chart is displayed:

- **Show By**  
Choose the basis for the primary axis of the chart:
  - **Count**  
Number of files in each category
  - **File Size**  
Disk storage space occupied by each category
- **Show Container Heat By - Container reports only**
  - **Total Matches**  
Use a heat map based on the number of matching files in each container (normalized against the highest number of matches in any of the displayed containers).
  - **Percentage**  
Use a heat map based on the percentage of files in each container that match the filter criteria.
- **Container Chart Depth - Container reports only**  
Set the number of levels of sub-folders to be displayed in Container reports. The value is retained for future reporting sessions.
- **Chart Axes**  
Choose the order of categories on the X-axis of the graph and filter results by File Count:
  - **File Count**  
Filter results by applying an upper and/or lower limit to the number of files in each category. For example, to exclude all categories with less than 5 files from the chart, set the 'To:' value to 5.
- **Order By**  
Order the categories in the chart
  - **Axis label**  
Display categories in alphabetical order
  - **Aggregate size**  
Display categories in decreasing Y-value
- **Include 'No Value' Results**  
(*Calculated Fields report only*)  
By default, Discovery Center excludes files from the report that do not match the chosen indexing criteria. This is to avoid the chart and file data being dominated by 'No value' results. Select this option if you want to include 'No value' files in the report.

## Data Table tab

The *Data Table* tab presents report data in tabulated form showing all statistics, regardless of any settings previously selected on the *Chart* tab.

## Sorting

Click on the column headers to reorder the listed data. Click on the column header a second time to reverse the sort order.





## Data Table Actions

You can use commands on the *Actions* menu to delete, migrate or markup the files in selected categories. Click on the check box alongside each category to select it. Select the check box at the top of the list to select all files on the current report view. Right click on the *Data Table* or click on the *Actions* button to view the available actions.

<input type="checkbox"/>	Folder Path	Folder Name	File Count	Container Count	Size	Matched File Co	Matched Size
<input type="checkbox"/>	\\localhost\test old docs\	test old docs	0	12	0 B	0	0 B
<input type="checkbox"/>	\\localhost\test old docs\deep folder structure\	deep folder structure	1	1	316.60 KB	1	316.60 KB
<input type="checkbox"/>	\\localhost\test old docs\deep folder structure\very long fol very long folder and file names can caus	1	0	316.60 KB	1	316.60 KB	
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test Data\	Dupe Report Test Data	36	16	11.32 MB	33	5.98 MB
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test Data\All Exact Ma All Exact Matches		7	3	1.96 MB	7	1.96 MB
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test Data\All Exact Ma Folder 1		4	0	1.37 MB	4	1.37 MB
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test Data\All Exact Ma Folder 2		3	0	598.34 KB	3	598.34 KB
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test Data\All Exact Ma Folder 3		0	0	0 B	0	0 B
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test Data\All Near Mat All Near Matches		8	3	3.35 MB	6	121.50 KB
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test Data\All Near Mat Folder 1		2	0	50.50 KB	2	50.50 KB
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test Data\All Near Mat Folder 2		3	0	2.13 MB	2	33.50 KB
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test Data\All Near Mat Folder 3		3	0	1.17 MB	2	37.50 KB
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test Data\Mixed Matc Mixed Matches No Overlap		10	3	4.45 MB	9	2.35 MB

Figure 79 Data Table tab

## Container List tab

The Container List tab shows the locations of files which match the report chart or data table views' selections, subject to any filters applied in defining the report.

### Sorting

Click on the column headers to reorder the listed data. Click on the column header a second time to reverse the sort order.

### Container List Page Sizes

The Container List tab presents files in pages according to the controls in the table footer. By default, each page lists up to 100 files although you can change this setting to other values using the dropdown control. Use the other controls to browse through additional pages. Alternatively, type the page number you want to display.



## Container List Actions

Right click on the *Container List* tab or click on the **Actions** button to view the available actions. You can use commands on the *Actions* menu to delete, migrate or mark up folders selected on the *Container List*. The **Export Containers for selections to CSV** command allows you to export a list of the folders and associated data. This applies only to selections on the *Chart* or *Data Table* tabs (not on the *Container List* itself). When the Container report has a heat overlay, you will be prompted to choose whether to include all folders in the export list or just those matching the filter selection.

[Back](#) | [Define View](#) | [Chart Settings](#) ▾ | [Save View](#) | [Actions](#) ▾

Title: Containers by Count on AOI: TOD  
Filter: Generic IM Policy ROT has any value.

Chart | Data Table | **Container List** | File List

<input type="checkbox"/>	Parent Path	Folder Name	File Count	% Files by Count	Size ↕	% Files by Size
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test [	Folder 2	5 (of 5)	100%	3.28 MB	100%
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test [	Folder 2	3 (of 3)	100%	2.13 MB	100%
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test [	Folder 1	4 (of 4)	100%	1.37 MB	100%
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test [	Folder 3	3 (of 3)	100%	1.17 MB	100%
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test [	Folder 3	3 (of 3)	100%	1.14 MB	100%
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test [	Folder 2	4 (of 4)	100%	623.38 KB	100%
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test [	Folder 1	4 (of 4)	100%	617.88 KB	100%
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test [	Folder 2	3 (of 3)	100%	598.34 KB	100%
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test [	Folder 3	3 (of 3)	100%	351.50 KB	100%
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test [	Folder 1	2 (of 2)	100%	50.50 KB	100%
<input type="checkbox"/>	\\localhost\test old docs\Dupe Report Test [	Folder 1	2 (of 2)	100%	34.00 KB	100%

Page 1 of 1 50 ▾

Figure 80 Container List tab

## File List tab

The File List provides access to the detailed results of indexing allowing you to inspect file metadata and carry out delete, migrate or markup actions on a file-by-file basis. Click on a file name hyperlink to display its basic properties, themes and summaries and calculated fields.

Index security credentials determine, based upon Windows Authentication, if you are able to view the results and contents of any given index. Index security settings have no effect on charts or data tables. However, those settings restrict access to file level reporting such as CSV exports, the File List and its file metadata views.

If you attempt to view file level information for an index to which you have no rights, the results will be filtered to remove those files from the relevant report. This may result in no data being shown.

To view the file list:



- For all the files in the report, click on the **File List** tab with none or all of the chart columns or slices selected.
- For specific file categories, select one or more bars or pie-slices on a report and then click on the **File List** tab.

The example shown in Figure 81 shows a basic File List. It contains extra information for the following report types:

- A Calculated Fields report (see page 128) includes a column for the current field value, *diversity* and *intensity*.
- The Duplicate reports grid (see page 133) features additional *Duplicate Status* and *Id* columns and groups matching files if the grid is sorted by Duplicate Id.
- The Containers report (see page 130) features a column for the value of the selected Heat field (if applicable).

Back | [Define View](#) | [Chart Settings](#) | [Save View](#) | [Actions](#)

Title: Containers by Count on \\localhost\test old docs\  
 Current View: Containers by Count  
 Filter: file size smaller than 1 MB.

Chart | Data Table | Container List | **File List**

<input type="checkbox"/>	Filename	Folder Path	Extension	Size	Created Date	Modified Date	Last Accessed Date	Owner
<input type="checkbox"/>	1a Australasia Profiles .ppt	\\localhost\test old docs\near_identical_duplicate.ppt	.ppt	860.50 KB	2014/06/08 19:07	2002/08/13 09:45	2014/06/08 19:07	REVIEW\techauthor
<input type="checkbox"/>	1a Australasia Profiles .ppt	\\localhost\test old docs\temp\near_identical_duplicate.ppt	.ppt	860.50 KB	2014/06/08 19:07	2002/08/13 09:45	2014/06/08 19:07	REVIEW\techauthor
<input type="checkbox"/>	Exact Copy of 1a Australasia Profiles .ppt	\\localhost\test old docs\near_identical_duplicate.ppt	.ppt	860.50 KB	2014/06/08 19:07	2002/08/13 09:45	2014/06/08 19:08	REVIEW\techauthor
<input type="checkbox"/>	Exact Copy of 1a Australasia Profiles .ppt	\\localhost\test old docs\temp\near_identical_duplicate.ppt	.ppt	860.50 KB	2014/06/08 19:07	2002/08/13 09:45	2014/06/08 19:08	REVIEW\techauthor
<input type="checkbox"/>	AN Taxonomy Editor Guide.doc	\\localhost\test old docs\Migrate Files\duplicate.doc	.doc	839.00 KB	2014/06/08 19:07	2007/08/08 14:50	2014/06/08 19:07	REVIEW\techauthor
<input type="checkbox"/>	AN Client Applications Installatio.doc	\\localhost\test old docs\Migrate Files\duplicate.doc	.doc	825.50 KB	2014/06/08 19:07	2007/09/17 14:56	2014/06/08 19:07	REVIEW\techauthor
<input type="checkbox"/>	WhitePaper £ Technical V3.doc	\\localhost\test old docs\unusual characters .doc	.doc	682.00 KB	2014/06/08 19:07	2005/04/04 16:08	2014/06/08 19:07	REVIEW\techauthor
<input type="checkbox"/>	Content Audit Checklist.doc	\\localhost\test old docs\Migrate Files\duplicate.doc	.doc	659.50 KB	2014/06/08 19:07	2007/09/17 10:45	2014/06/08 19:07	REVIEW\techauthor
<input type="checkbox"/>	cs_ct01.pdf	\\localhost\test old docs\Dupe Report Test Data\ .pdf	.pdf	581.80 KB	2014/06/08 19:07	2005/05/18 22:37	2014/06/08 19:08	REVIEW\techauthor
<input type="checkbox"/>	cs_ct01.pdf	\\localhost\test old docs\Dupe Report Test Data\ .pdf	.pdf	581.80 KB	2014/06/08 19:07	2005/05/18 22:37	2014/06/08 19:08	REVIEW\techauthor
<input type="checkbox"/>	cs_ct01.pdf	\\localhost\test old docs\Dupe Report Test Data\ .pdf	.pdf	581.80 KB	2014/06/08 19:07	2005/05/18 22:37	2014/06/08 19:08	REVIEW\techauthor
<input type="checkbox"/>	cs_ct01.pdf	\\localhost\test old docs\Dupe Report Test Data\ .pdf	.pdf	581.80 KB	2014/06/08 19:07	2005/05/18 22:37	2014/06/08 19:08	REVIEW\techauthor
<input type="checkbox"/>	cs_ct01.pdf	\\localhost\test old docs\Unactioned Dupes Test .pdf	.pdf	581.80 KB	2014/06/08 19:07	2005/05/18 22:37	2014/06/08 19:08	REVIEW\techauthor
<input type="checkbox"/>	cs_ct01.pdf	\\localhost\test old docs\Unactioned Dupes Test .pdf	.pdf	581.80 KB	2014/06/08 19:07	2005/05/18 22:37	2014/06/08 19:08	REVIEW\techauthor
<input type="checkbox"/>	cs_ct01.pdf	\\localhost\test old docs\Unactioned Dupes Test .pdf	.pdf	581.80 KB	2014/06/08 19:07	2005/05/18 22:37	2014/06/08 19:08	REVIEW\techauthor
<input type="checkbox"/>	cs_ct01.pdf	\\localhost\test old docs\Unactioned Dupes Test .pdf	.pdf	581.80 KB	2014/06/08 19:07	2005/05/18 22:37	2014/06/08 19:08	REVIEW\techauthor
<input type="checkbox"/>	cs_ct01.pdf	\\localhost\test old docs\Unactioned Dupes Test .pdf	.pdf	581.80 KB	2014/06/08 19:07	2005/05/18 22:37	2014/06/08 19:08	REVIEW\techauthor

Figure 81 Example file list

### Sorting

Click on the column headers to reorder the listed files. Click on the column header a second time to reverse the sort order.

### File List Page Sizes

The *File List* tab presents files in pages according to the controls in the table footer. By default, each page lists up to 50 files although you can change this setting to 25 or 100 using the dropdown control. Use the other controls to browse through additional pages. Alternatively, type the page number you want to display.

### File List Actions

You can use commands on the *Actions* menu to delete, migrate or markup selected files. Click on the check box alongside each file to select it. Select the check box at the top of the list to select all files on the currently displayed page. Right click on the file list or click on the **Actions** button to view the available actions.



## Exporting a File List

Select the **Export current file list** command on the *Actions* menu to download the full set of files for the selected chart elements (bars or pie slices). The *Select CSV Export Columns* dialog box is displayed, offering options to choose which Basic Metadata and Calculated Fields are included in the export file (see page 153).

Report exports can provide full details, including all metadata and properties, of every file in a chart or table selection, exported to a CSV format. This format avoids the limitations inherent in 3rd-party applications (such as Microsoft Excel). You may generate an export containing many millions of files; these exports will take some time to produce and download and will require space on disk whilst they are built.

## File Metadata Preview

Click on the name of a file in the *File List* to view the detailed results of indexing in the *File Metadata Preview* window. The window displays a link to the original file; click on this to review the source file (if you have permission to access it and an application to view the file).

**Note.** The ability to open file system links will be dependent on your browser. To enable opening of such links in Firefox you must enable specific settings as described [http://kb.mozillazine.org/Links\\_to\\_local\\_pages\\_don't\\_work](http://kb.mozillazine.org/Links_to_local_pages_don't_work).

The file properties are shown in four tabbed pages:

- Basic Metadata
- Themes & Summaries
- Calculated Fields
- Markup

If metadata is not available for the selected file, the corresponding tab will be grayed out. A link at the top of the window allows you to access the selected file. If you are reviewing many files, use the **Next** and **Previous** buttons to step through the File List.



## Basic Metadata

The *Basic Metadata* tab shows the selected file's name, folder and file path, size, format and the created, modified and last accessed dates. If available, the tab also shows the file owner property. This is optionally collected during the skim from the file owner property in Windows file shares and the created by property from files in SharePoint.

**Note.** Retrieving the file ownership property can reduce skim performance (depending on the state of the local environment) and can be disabled in Advanced Index Options.

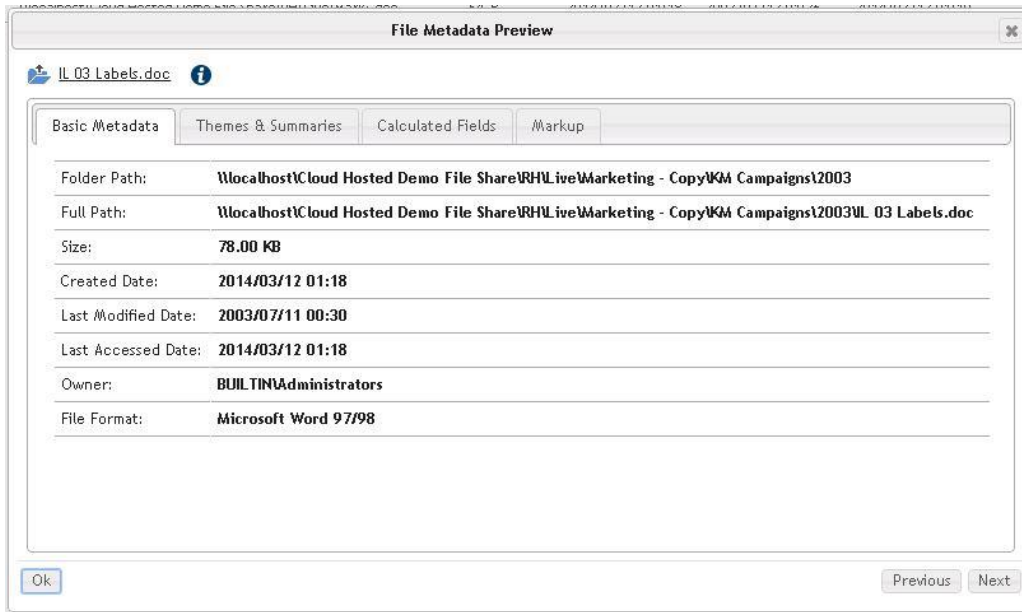


Figure 82 File Metadata Preview: Basic Metadata tab

## Themes and Summaries tab

Themes are shown with a colored background according to their relative weight in the file. Higher weighted themes will be darker with larger text than lower weighted themes. Summaries are shown as summary sentences, each separated with an ellipsis.

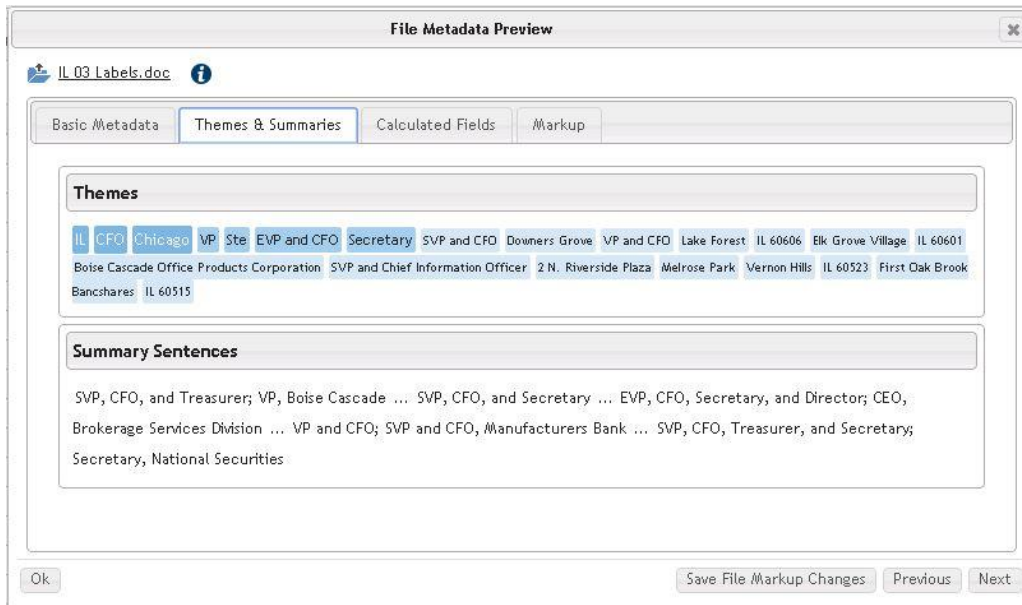


Figure 83 Themes & Summaries tab

## Calculated Fields tab

All indexed calculated fields are listed with their values and intensity and diversity scores.

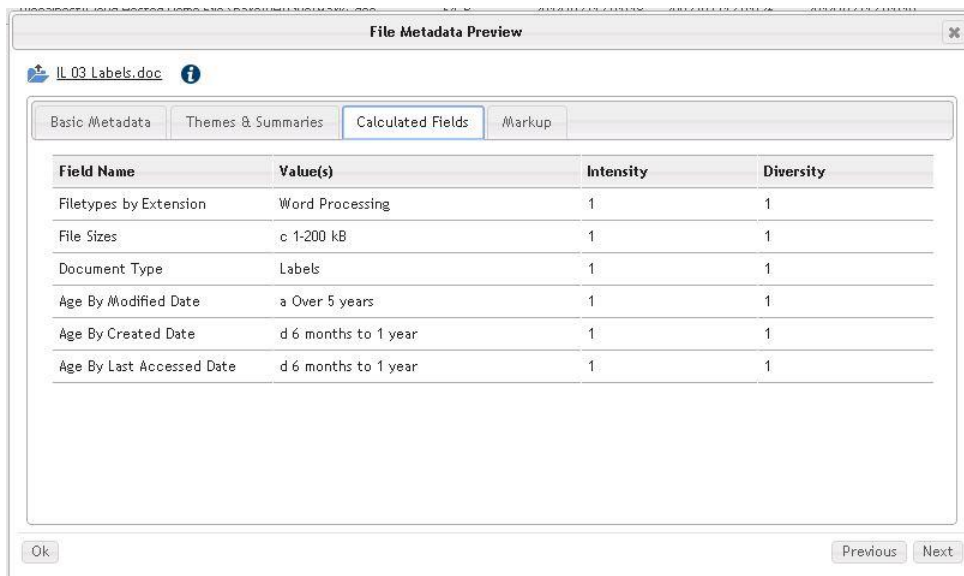
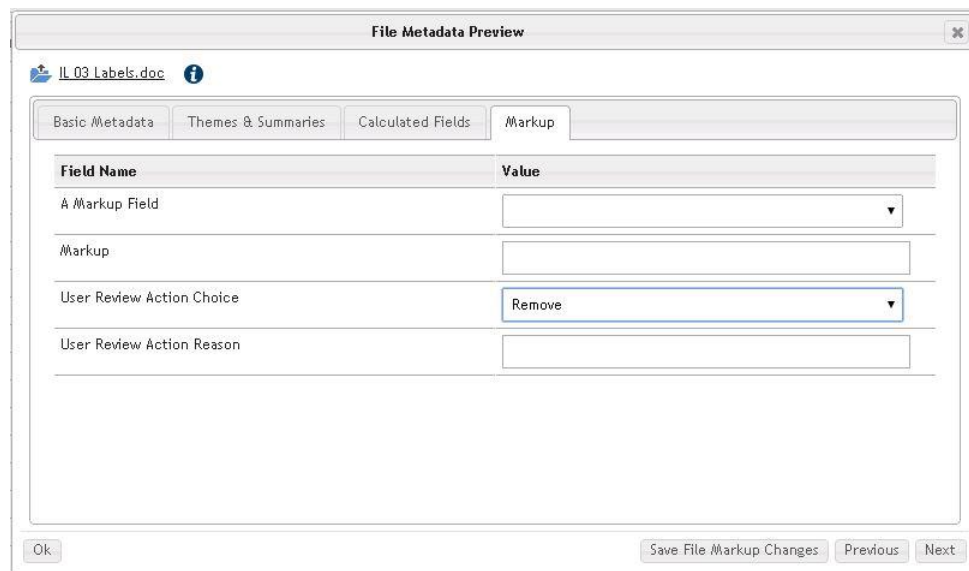


Figure 84 Calculated Fields tab

## Markup

Markup fields are listed with their current values. To change the file's markup, enter text, or pick from the list of allowed choices (depending on the options set by the AN Administrator, see page 59). If you have changed any of the markup fields, you will be prompted to save your changes: click on the **Save File Markup Changes** to continue.



The screenshot shows a window titled "File Metadata Preview" for the file "IL\_03\_Labels.doc". It has four tabs: "Basic Metadata", "Themes & Summaries", "Calculated Fields", and "Markup". The "Markup" tab is active, displaying a table with two columns: "Field Name" and "Value".

Field Name	Value
A Markup Field	<input type="text"/>
Markup	<input type="text"/>
User Review Action Choice	<input type="text" value="Remove"/>
User Review Action Reason	<input type="text"/>

At the bottom of the dialog, there are buttons for "Ok", "Save File Markup Changes", "Previous", and "Next".

Figure 85 Markup tab

**Note.** Each time you save metadata for a file, a new action will be added to the activity queue. If you need to update metadata in bulk, then consider using the **Markup** action from the *Chart*, *Data Table* or use multiple selections from the *File List*.

## Actions

Depending on the Discovery Center features that are licensed on your system, commands on the *Actions* menu allow Information Managers to:

- Export Data Table
- Export Container List
- Export File List
- Delete
- Quarantine
- Migrate
- Update
- Markup
- Set View Location (Containers report only)
- Select all files
- Select all 100% match locations (Containers report with heat overlay only)
- Deselect all files

**Note.** When a report is opened in Review mode from the *Work Packages* tab, *Markup Selected Fields* is the only supported action.

For example, having generated a report showing the distribution of file ages, you could delete all files created before a specific year, or having used metadata to identify files across the network that contain sensitive information, you could migrate the files to a single secure location.

To carry out actions on files, containers or data in a report:

1. If the relevant report is not currently displayed in Discovery Center, go to the Reporting and Actions page, locate and click on the report in the Saved Views list.
2. Select the files from the Chart, Data Table or File List Views:
  - On the chart, click on the categories or chart segments. Selected chart elements are highlighted with a red outline. To select/deselect all files, click on the **Actions** button and choose the *Select/Deselect all files* command.
  - In the *Data Table* or *File List*, click on the check box alongside each file. Select the check box at the top of the list to select all files.
  - In the *File List* or *Container List*, export actions are applied to selections on the **Chart** or **Data Table** tabs.
3. Right click on the chart or click on the **Actions** link. Choose the required action as described in the following sections.

[Back](#) | [Define View](#) | [Chart Settings](#) | [Save View](#) | [Actions](#)

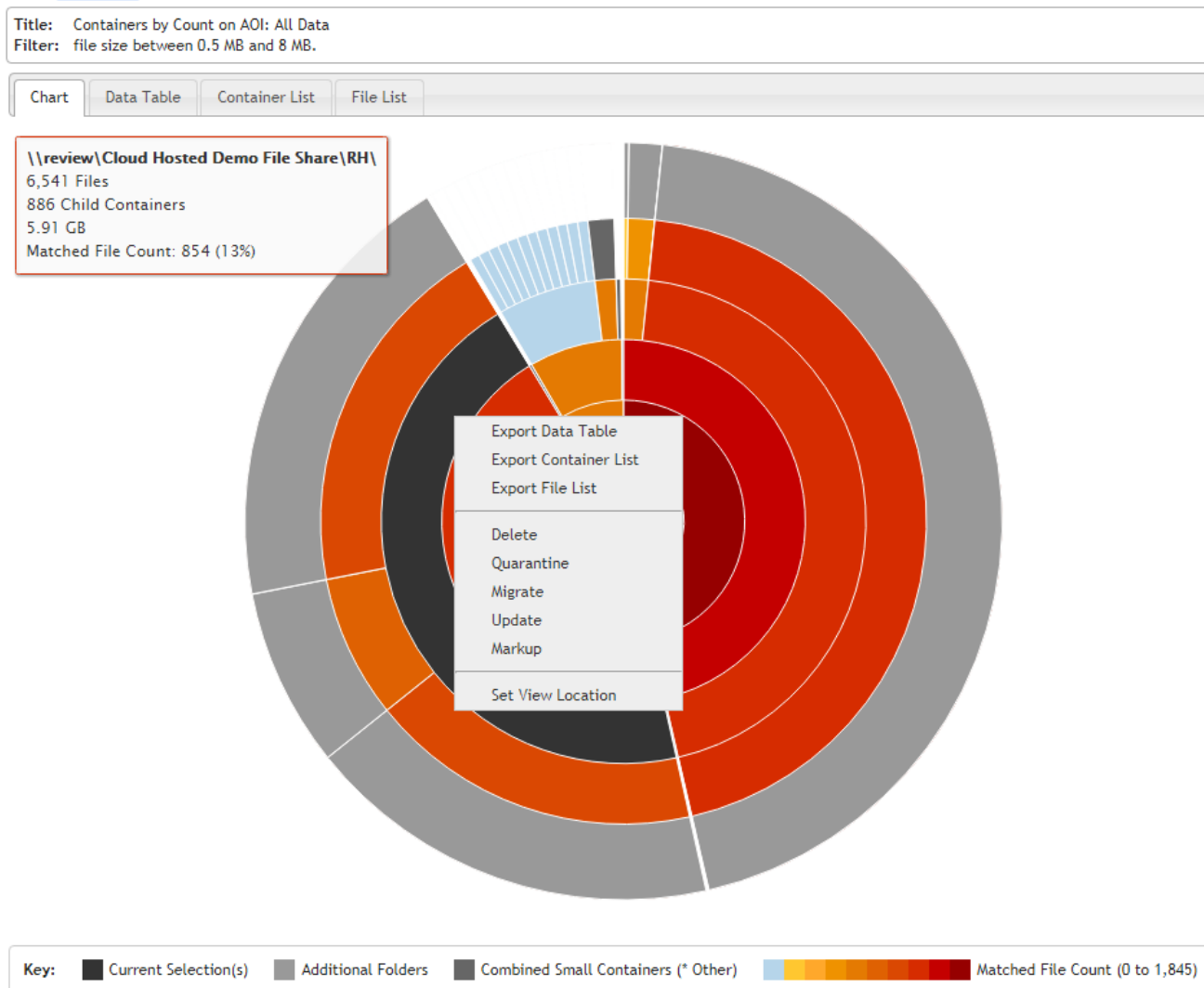


Figure 86 The Actions menu





## Export Data Table

Save or open the report information presented on the Data Table tab in CSV format.

## Export Container List

The **Export Container List** command allows you to export a list of the folders in a Container report. The export includes detailed information including Parent Path, Folder Name, Matched Files, Total Files, Matched Size, Total Size, Matched Files Percentage, Matched Size Percentage. It applies only to selections on the *Chart* or *Data Table* tabs (not on the *Container List* or *File List* tabs). When the Container report has a heat overlay, you will be prompted to choose whether to include all folders in the export list or just those matching the filter selection.

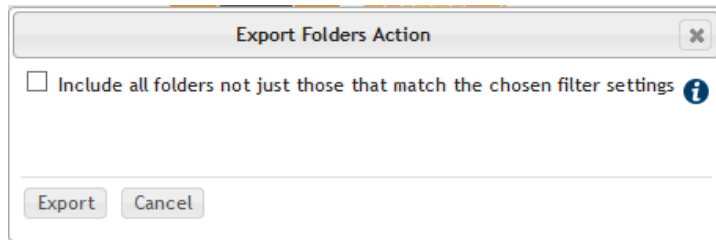


Figure 87 Export Folders Action

## Export File List

The **Export File List** command exports a file list in either CSV or XLSX format (determined by **Export File Type** setting on Reporting Settings page). It is restricted to files matching selections in the current Report View tab. If you are viewing the *File List* tab and have not selected any data table elements the complete file list is available for download.

**Note.** It is easy to generate an export containing many millions of files; these exports will take some time to produce and download and will require space on disk while they are built. CSV format avoids the limitations inherent in 3rd-party applications (such as Microsoft Excel).

When the **Export File List** command is selected the *Export Action* dialog box is displayed, offering options to choose which Basic Metadata and Calculated Fields are included in the export file. Use this feature to reduce the size of export files and save time manually tidying exported files during subsequent data analysis.

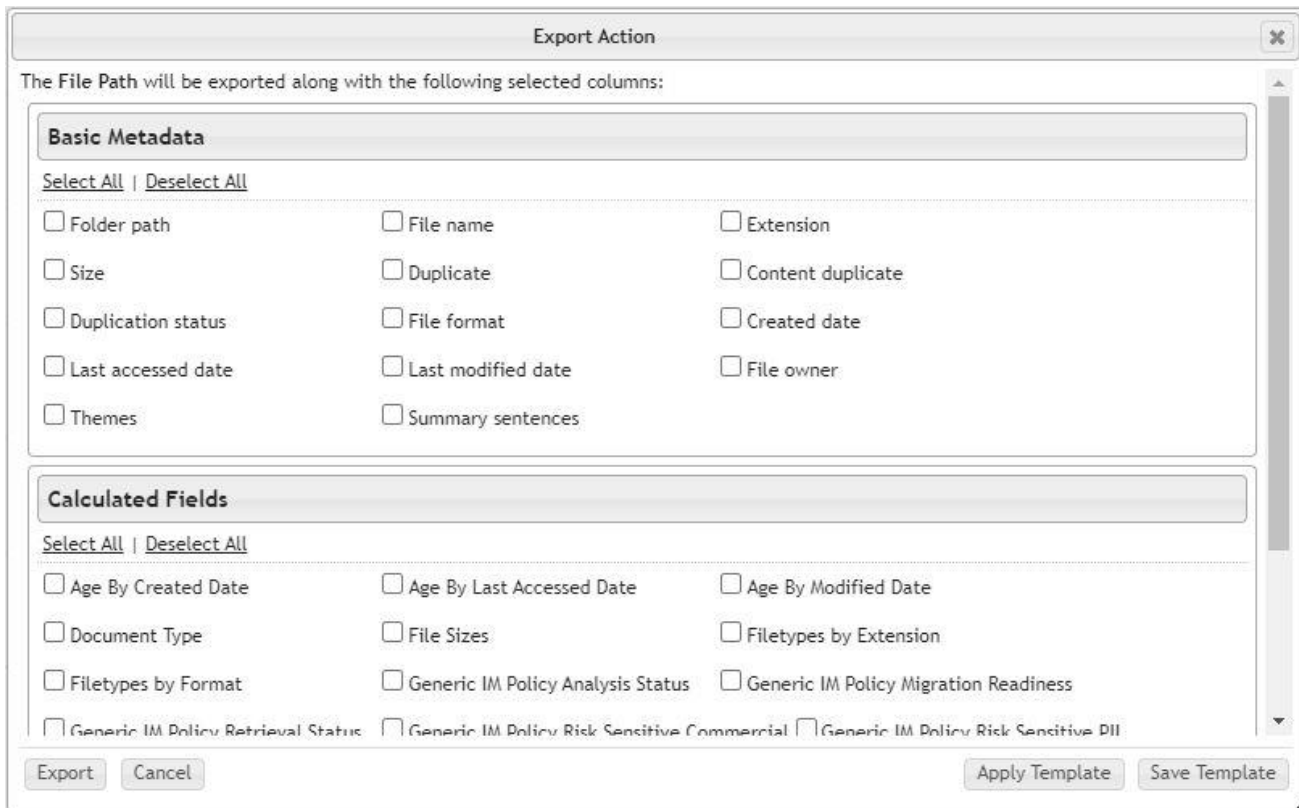


Figure 88 Select File Export Columns

*Containers reports with heat only*

**Include all files in folder not just those that match the chosen filter settings**

By default, the Export action is only applied to files that match the filter selection applied to the report. Select this option if you want to export information for all files in a folder, even if they do not match the filter settings.

You can also choose whether or not to include Intensity/Diversity scores. The **Field Intensity** score is the total number of hits found by the extraction rules (or classification). Intensity is a gross indicator of the rule success. **Field Diversity** measures the number of different values being found: a high score indicates multiple hits with different values. Conversely, multiple hits with the same value will attract a high Intensity count but a low Diversity value. For sensitive data analysis, the Diversity score shows how risky the file might be.

The **Save Template** button allows you to save the current selections made in the checkboxes in the dialog box so they can be re-applied for future file list exports using the **Apply Template** button, which will populate the checkboxes in the dialog box with the previously saved selections. There can only be one saved selection template for the application shared between all users, and when the **Save Template** option is chosen any previous saved selections will be overwritten by the current selections. By default, the saved template is an empty configuration - i.e. all checkboxes will be cleared when **Apply Template** is selected. Once a template has been applied the selections can still be changed prior to the files being exported, the configuration used in the export will always reflect the selections made in the checkboxes at the point that the **Export** button is selected.

When the *Export Action* dialog box is first loaded it may include some selections by default. When this occurs, it is retrieving the configuration used for the previous file list export from a browser cookie, so this configuration only applies for the current user and the current browser in use. This configuration can be amended by making different selections or overridden using the **Apply Template** button if required. When a file list export is performed the selections used will be persisted into the browser cookie again.



## Delete

**Note.** If you are working with a Duplicates report, the *Delete* action is only permitted on duplicate files (the *File Duplicate* or *Content Duplicate* elements of the analysis). The action is not permitted if you have selected one or more of the *Master*, *Not Duplicated*, *Unanalyzed* or *Empty* elements.

Select the files as described above and then choose **Delete** from the Actions menu. Choose from the following options:

- *Containers report with heat only*  
**Include all files in folder not just those that match the chosen filter settings**  
The Delete action only removes files that match the filter selection applied to the report. Select this option if you want to remove all files from the folder, even if they do not match the filter settings.
- *Duplicates report only*  
**Create shortcuts to the master document**  
Delete the selected files but create, in their place, shortcuts to the corresponding master files. In each cluster of duplicate files, the 'master' file has been identified using the specified selection strategy: First Created, Last Modified, or by Document Location/AOI.
- **Cleanup empty folders during deletion**  
Select this option if you want to delete any empty containers that may be created as a result of the delete action. This will not affect any empty folders that already existed before the action.
- **Action Annotation**  
Add markup text describing the reasons for this action (optional).

Click on **Delete** to confirm the selected options and queue the actions for processing.

## Quarantine

**Note.** If you are working with a Duplicates report, the *Quarantine* action is only permitted on duplicate files (the *File Duplicate* or *Content Duplicate* elements of the analysis). The action is not permitted if you have selected one or more of the *Master*, *Not Duplicated*, *Unanalyzed* or *Empty* elements.

Select the files as described above and then choose Quarantine files from the Actions menu. Choose from the following options:

- *Containers reports with heat only*  
**Include all files in container not just those that match the chosen filter settings**  
The Quarantine action only moves files that match the filter selection applied to the report. Select this option if you want to move all files from the folder, even if they do not match the filter settings.
- **Quarantine Location**  
Select the required quarantine location from the dropdown list (as specified on the Reporting Settings tab, see page 176).  
**Note.** The quarantine location will require careful management to ensure that files can be restored, if necessary, and all genuine ROT is discarded after a suitable time period. If the quarantine location is included in an index, its contents will contribute to the advisory Volume Under Management total (if you have a limited volume license).
- *Duplicates reports only*  
**Create shortcuts to the master document**  
Quarantine the selected files but create, in their place, shortcuts to the corresponding master files. In each cluster of duplicate files, the 'master' file has been identified using the specified selection strategy: First Created, Last Modified, or by Document Location/AOI.
- Cleanup empty containers during quarantine
  - **Take no action**  
Retain empty containers.



- **Remove all**  
Remove all empty containers from the selection, including those that are created as a result of the action.
- **Remove containers emptied by this action**  
Only remove empty containers that are created as a result of the action. This will not affect any empty folders that already existed before the action.
- **Metadata Mapping Set**  
Use a Metadata Mapping Set (see page 164) to ensure that important discovered and calculated information is added to the quarantine location as metadata. Select the required Mapping Set from the dropdown list (visit the Mapping Rules tab to create a new Mapping Set if necessary) or choose the No mapping set option if you do not want to carry out metadata mapping during the migration.
- **Character Mapping Set**  
Use a Character Mapping Set (see page 169) to ensure that illegal characters are removed or replaced with acceptable text or other characters before moving to the quarantine location. Select the required Mapping Set from the dropdown list (visit the Mapping Rules tab to create a new Mapping Set if necessary) or choose the No character mapping set option if you do not want to carry out character mapping.
- **Action Annotation**  
Add markup text describing the reasons for this action (optional).

Click on **Quarantine** to confirm the selected options and queue the actions for processing.

## Migrate

Select the files as described above and then choose **Migrate** from the *Actions* menu. You are prompted to select the following migration options:

- *Containers report with heat only*  
**Include all files in folder not just those that match the chosen filter settings**  
The Migrate action only moves files that match the filter selection applied to the report. Select this option if you want to migrate all files from the folder, even if they do not match the filter settings.

### Destination Folder

Identify the destination for the files to be migrated.

### Destination Folder Structure

Choose one of the following options:

- **No Folders**  
Copy all the files to the selected folder without creating a folder structure.
- **Replicate Source Folders**  
Reproduce the folder structure of the selected files at the new location.
  - **Number of folder levels**  
Specify the number of folder levels to be recreated in the new location. If you choose to include the parent folder (see below) this is included as the first 'level'. Otherwise, its child folders are considered to be the first level folders. Select **All** to reproduce the entire folder hierarchy (with or without the parent folder) in the new location.
  - **Include parent folder**  
Choose whether to reproduce the root folder in the new location.
  - **Copy Source Folder Permissions**  
Reproduce the folder permissions in the new, destination folder structure.



The folder is initially created with any permissions inherited from its parent. Discovery Center also merges any missing permissions from the source location so no effective access is lost relative to the source. This may entail the creation of apparently overlapping permissions for a given user in the destination folder, dependent on the way in which permissions may have been inherited for that user from the parent hierarchy.

If the Destination folders already exist, choose from one of the following three options to determine how to address any conflicts over access permissions:

- **Use Existing Permissions**  
Do not update the permissions for destination folders which already exist.
- **Merge Permissions**  
Keep the existing permissions for each destination folder and add those from the corresponding source folder.
- **Abort**  
Do not migrate files to pre-existing destination folders.

**Note.** When copying permissions, BUILTIN groups for the local system will be copied across (for example, BUILTIN\Administrators and BUILTIN\Users). When an action moves files between two locations, these groups may have different members in source and destination locations.

When folders are encountered that explicitly break the inheritance of parent permissions, the *Copy Source Folder Permissions* option will respect this permission pattern. However, if the change in inheritance affects the access to folders for the accounts used by the ActiveNav Discovery Center, some files may not be migrated successfully.

Table 14 How migration settings affect folder permissions

Migration Settings		Folder exists	Folder is new
Copy Source Folder Permissions?	No	Maintain existing destination folder permissions	Inherit parent folder permissions only
	Use Existing Permissions	Maintain existing destination folder permissions	Inherit parent folder permissions and merge with any additional permissions from source folder
	Yes Merge Permissions	Keep the existing destination folder permissions and add those from the corresponding source folder	
	Abort	Skip the migration of files into any pre-existing destination folders	

- **Calculated Field Values**

Create a new structure based on the file classification.

- **Migration Field**  
Choose the Calculated Field that will form the basis of the file migration. The values of the Calculated Field will form the folder structure for the migrated files with each file assigned to folder depending on its own value.
- **Migrate unclassified files**  
By default, files that do not have a value for the selected Calculated Field will be migrated to a folder named "AN Unclassified" in the destination folder structure. If you do not want to migrate unclassified files, clear the *Migrate Unclassified fields* check box. The SharePoint connector can create document libraries and folders where appropriate (when they do not exist) but it will not create new sites.



## Original File

Choose whether to retain or delete the original files from the source location, or to replace each file with a shortcut linking to the appropriate migrated file in the destination folder.

- **Keep**  
Leave the original files in the source location.
- **Leave Shortcut**  
Replace each file in the source folder with a shortcut linking to the appropriate migrated file in the destination folder. Discovery Center cannot create shortcuts in SharePoint repositories, but can create shortcuts to SharePoint.
- **Remove**  
Delete the original file from the source folder (default).
- **Cleanup empty folders during migration**  
Select this option if you want to delete any empty containers that may be created as a result of an action that is removing files. This will not remove any empty folders that already existed before the action.

## Sensitivity Label Mapping Source Field

If your license allows it you can use this option to select a Calculated Field to use as the source of the value of an MIP Sensitivity Label to apply to any supported files as part of the migration process. The Calculated Fields that appear in this list are limited to those that can hold a single value only, as only one MIP Sensitivity Label can be applied to a given file at a time (See **Calculated Fields** for further details on how to configure Calculated Fields).

For instance, you may have a Single Value Classification Calculated Field named ‘Sensitivity’ defined, whose values each correspond to the value of an MIP Sensitivity Label available in your O365 instance. If the files being migrated have previously been indexed and classified so they have a value associated with the ‘Sensitivity’ field, and are a type that supports MIP Sensitivity Labels, then choosing the ‘Sensitivity’ field here would cause the value held in the field to be set as the MIP Sensitivity Label for the file as part of the migration.

For the application of a Sensitivity Label to be successful during migration the value held for the chosen Calculated Field for the file in question has to match exactly with a valid MIP Sensitivity Label in the configured O365 instance and the correct MIP Settings need to have been applied on the **Discovery Center System Settings** page. This process supports updating the MIP Sensitivity Label value that is set on a file during migration, but does not provide the means to remove any MIP Sensitivity Label that has previously been applied.

## Allow MIP Sensitivity Label Downgrade

If a **Sensitivity Label Mapping Source Field** has been chosen this option allows you to choose whether the downgrade of MIP Sensitivity Labels is allowed in the cases where the file being migrated already has an MIP Sensitivity Label applied and value held in the chosen Calculated Field represents a label of a lower order of precedence in O365.

## MIP Sensitivity Label Downgrade Justification

If the **Allow MIP Sensitivity Label Downgrade** option is selected then a justification reason is required for the cases where the file migration process would result in a lower precedence MIP Sensitivity Label being applied to the destination file as compared with the original.



## Metadata Mapping Set

Use a Metadata Mapping Set (see page 164) to ensure that important discovered and calculated information is added to the new location as metadata. Select the required Mapping Set from the dropdown list or choose the *No mapping* set option if you do not want to carry out metadata mapping during the migration.

## Character Mapping Set

Use a Character Mapping Set (see page 169) to ensure that illegal characters are removed or replaced with acceptable text or other characters before migration to the new location. Select the required Mapping Set from the dropdown list or choose the *No character mapping* set option if you do not want to carry out character mapping during the migration.

## Action Annotation

Add markup text describing the reasons for this action (optional).

**Migrate Action**

This action will be performed on 6 selections including .xlsx, .docx, .pptx, .jpg, .txt, .png ...

**Destination Options**

Destination Container  
\\orgfileserver01\Secureshare\ Browse... *i*

Destination Container Structure  
Replicate Source Containers *i*

Number of Container Levels  
All *i*

Include Parent Container *i*

Copy Source Container Permissions *i*

If Destination Folder Already Exists  
 Use Existing Permissions  Merge Permissions  Abort *i*

**Source Options**

Original File  
Keep *i*

Cleanup Empty Source Containers  
Take no action *i*

**MIP Sensitivity Label Mapping Options**

MIP Sensitivity Label Mapping Source Field  
Markup *i*

Allow MIP Sensitivity Label Downgrade *i*

MIP Sensitivity Label Downgrade Justification  
*i*

**Mapping Options**

Metadata Mapping Set  
No mapping set *i*

Character Mapping Set  
No character mapping set *i*

Action Annotation

Migrate Cancel

Figure 89 Migrate Action

## Update Metadata

Use this option to update selected files in a repository such as SharePoint with metadata from the Discovery Center database.

1. *Containers report with heat only*

**Include all files in folder not just those that match the chosen filter settings**

The Update Metadata action only acts on files that match the filter selection applied to the report. Select this option if you want to update metadata for all files in the folder, even if they do not match the filter settings.

2. Select a Metadata Mapping Set.
3. Use the **Action Annotation** box to add further information about the markup (optional).
4. Click on the **Update Metadata** button.
5. Click on the **OK** button to confirm that you want to update the metadata of the selected files using the values derived by Discovery Center and the mapping defined by the selected mapping set.
6. The update action is added to the processing queue.

## Markup

This option allows you to mark up all the files selected in the chart with a custom field value. You could use this to identify files for review, for example. To be able to use this option, the AN Administrator must first create one or more Markup fields, which are not assigned a value during analysis, or have made a standard calculated field available for markup.

1. *Containers report with heat only*

**Include all files in folder not just those that match the chosen filter settings**

The Markup action only acts on files that match the filter selection applied to the report. Select this option if you want to mark up all files in the folder, even if they do not match the filter settings.

2. Select the Markup field.

If you choose a calculated field that has been made available for markup the following warning is displayed:

*“The field selected for markup is not a markup field so file field values will be overwritten when a parent index is re-processed.”*

3. Enter the Markup value.

You will be able to enter text, or pick from a list of allowed choices depending on the options set by the AN Administrator.

4. Use the **Action Annotation** box to add further information about the markup (optional).
5. Click on the **Markup** button.

You can examine files with Markup field values using a *Calculated Fields* report.

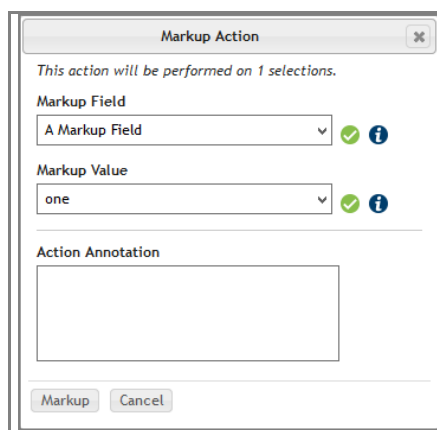


Figure 90 Markup Action



## MIP Sensitivity Label (In Place)

If your license allows it you can use this option to apply a MIP Sensitivity Label to any supported files selected. This will be an in-place action i.e. files will be labeled in their current location.

Select the files to be labeled and then choose **MIP Sensitivity Label** from the Actions menu. You are prompted to select the following options:

### 1. MIP Sensitivity Label Mapping Source Field

Use this option to select a Calculated Field to use as the source of the value of an MIP Sensitivity Label to apply to any supported files. The Calculated Fields that appear in this list are limited to those that can hold a single value only, as only one MIP Sensitivity Label can be applied to a given file at a time (See [Calculated Fields](#) for further details on how to configure Calculated Fields).

For instance, you may have a Single Value Classification Calculated Field named 'Sensitivity' defined, whose values each correspond to the value of an MIP Sensitivity Label available in your O365 instance. If the selected files have previously been indexed and classified so they have a value associated with the 'Sensitivity' field, and are a type that supports MIP Sensitivity Labels, then choosing the 'Sensitivity' field here would cause the value held in the field to be set as the MIP Sensitivity Label for the file.

For the application of a Sensitivity Label to be successful the value held for the chosen Calculated Field for the file in question has to match exactly with a valid MIP Sensitivity Label in the configured O365 instance and the correct MIP Settings need to have been applied on the [Discovery Center System Settings](#) page. This process supports updating the MIP Sensitivity Label value that is set on a file, but does not provide the means to remove any MIP Sensitivity Label that has previously been applied.

### 2. Encrypted File Extension Change Behavior

Use this option to specify what will happen to the original file when a selected label applies protection causing a change to the file extension. For example, an original file called "MyFile.txt" will result in a labeled and protected file called "MyFile.ptxt". Selecting Preserve Original means that the original file will be kept, Remove Original means that the original file will be deleted (and therefore only the file with the new extension will remain).

### 3. Allow MIP Sensitivity Label Downgrade

This option allows you to choose whether the downgrade of MIP Sensitivity Labels is allowed in the cases where the selected file already has an MIP Sensitivity Label applied and value held in the chosen Calculated Field represents a label of a lower order of precedence in O365.

### 4. MIP Sensitivity Label Downgrade Justification

The MIP label policy may require justification text when a file's sensitivity label is downgraded. If the Allow MIP Sensitivity Label Downgrade option is selected then MIP Sensitivity Label Downgrade Justification text is used to satisfy the MIP label policy when the MIP Sensitivity Label being applied to a file is of lower precedence than the original.

### 5. Action Annotation



This optional text is added to inform Administrators of the reasons for this action. It can be viewed in the Task Status under Activity History.

✕

MIP Sensitivity Label Action

*This action will be performed on the following selections: .docx*

---

**MIP Sensitivity Label Mapping Options**

MIP Sensitivity Label Mapping Source Field: Markup ✔ i

Encrypted File Extension Change Behavior: Preserve Original ✔ i

Allow MIP Sensitivity Label Downgrade i

MIP Sensitivity Label Downgrade Justification: Compliance review. ✔ i

---

Action Annotation:

---

Label Cancel

Figure 91 MIP Sensitivity Label Action

### File Extension Changes Following MIP Sensitivity Label Action

If the label applied to the file would result in a change to the file extension a new file will be created in the original file's current location and action taken on the original based on the Encrypted File Extension Change Behavior selected when running the action.

### Preserving File Metadata Following MIP Sensitivity Label Action

Where possible, Discovery Center will attempt to preserve file metadata following an MIP Sensitivity Label action. The deciding factors in what metadata can be preserved are:

- The repository where the files are held;
- Whether the labeling action resulted in a change to the file extension.
- The metadata retained is detailed in the below table.



Table 15 MIP Sensitivity Label Metadata Preservation

File System Metadata	Extension Not Updated	Extension Updated
<i>Created date</i>	Preserved	Preserved
<i>Last modified date</i>	Preserved	Preserved
<i>Last accessed date</i>	Preserved	Preserved
<i>NTFS File Owner</i>	Refer to Appendix 6 – Preserving File Owner	
SharePoint Metadata	Extension Not Updated	Extension Updated
<i>Created date</i>	Preserved	Preserved
<i>Modified Date</i>	Preserved	Preserved
<i>Created By</i>	Preserved	<b>SharePoint Online:</b> Set to SharePoint Connector Application. <b>SharePoint On-Premises:</b> Set to the credential used to perform the action.
<i>Modified By</i>	<b>SharePoint Online:</b> Set to SharePoint Connector Application. <b>SharePoint On-Premises:</b> Set to the credential used to perform the action.	<b>SharePoint Online:</b> Set to SharePoint Connector Application. <b>SharePoint On-Premises:</b> Set to the credential used to perform the action.
<i>Other SharePoint Metadata Values</i>	Preserved	Partially preserved (see below)

### Preserving SharePoint Metadata

When applying an MIP Label in SharePoint that results in the file extension being updated, a new file is created rather than an overwrite being performed. In this scenario any metadata values not detailed in Table 15 will be preserved in the new file, with the following exceptions:

- Metadata of type **User or Group** will not be preserved;
- Metadata of type **Hyperlink** will have the link value preserved, but not the description text;
- Metadata for the **Image Tags** column will not be preserved for image files within **Sharepoint Online**;



## Custom Queries

The Custom Query feature allows an AN Administrator to execute SQL queries directly against the Active Nav database.

(Custom Queries are disabled by default. To enable them, go to System Settings > Discovery Center > Global Settings and set Enable Custom Queries to true)

Several pre-installed Custom Queries are available in the Custom Queries folder under the install location, and new queries can be created and saved here.

Only custom queries that are in the install location and that conform to the specified format are available for execution.

When a Custom Query is run, it will return the results in a downloadable file to the executing user.

### Custom Query format

Custom Query files are in XML format and required to conform to a structure containing the following nodes:

- **CustomQuery Name**  
This is the name which is displayed in the dropdown on the Custom Queries page.
- **CustomQuery Description**  
This is descriptive text which is displayed when a custom query has been selected in the dropdown.
- **Query**  
This is where the desired SQL statements should be included; these are not displayed in the UI but will be run against the database when the custom query is executed.
- **Parameters**  
(optional) Parameters are values that the user can pass into the SQL statement. If a parameter is defined, the user must specify the value to be used in a textbox on the Custom Queries page.
- **Signature**  
Custom Queries must be signed in order to be executed in Discovery Center. Please contact ActiveNav Support to do this.



## Executing Custom Queries

To execute a query, select it from the dropdown on the Custom Queries page.

**Note.** Only valid, signed custom queries will be displayed in the dropdown list.

Discovery Center

Home Network Map System Settings Metadata Indexes Activity Reporting and Actions

Reporting Overview Saved Views Actions Work Packages Report Viewer Custom Queries Mapping Rules Reporting Settings

There were errors in 1 of the defined custom query files, please see the logs for more information.

Please select the custom query to execute:

List all failed containers for index

Lists containers that could not be retrieved for an index (RetrievalStatus >= 48)

Index name

Execute

CustomQuery Name  
Description  
Parameters

Figure 92 Custom Queries page

Enter any parameters required and click **Execute**.

The query is executed and results saved to a csv file which is downloaded.



# Mapping Rules

The Mapping Rules tab (available to users with the *AN Administrator* role only) has two sub-tabs:

- **Metadata Mappings**  
Match discovered and calculated information, based on calculated field values, to metadata values at the migration location (for example, SharePoint).
- **Character Mappings**  
Create a set of substitute text strings to replace illegal characters in file and folder names.

## Metadata Mappings

When an Information Manager chooses to migrate files as a reporting action, a metadata mapping set ensures that important discovered and calculated information is added to the new location as metadata. In addition, the Information Manager has the option to create a new hierarchy in the destination repository based on calculated field values.



Figure 93 Reporting and Actions page – Mapping Rules – Metadata Mappings

### Field Source

A **field source** is a location used to provide definitions of available metadata fields for mapping purposes. It may be the destination repository for file migration although this need not necessarily be the case so long as the contained metadata is equivalent. Before you can create a metadata mapping set, you must identify the field source to be used for metadata mapping (such as a SharePoint site collection). Discovery Center imports all of the metadata fields available at the selected location and makes these available for mapping against basic metadata and calculation fields discovered by an index.

### Adding a Mapping Set

1. To create a mapping set, click on the **Add Mapping Set** link. The *Define Mapping Set* dialog box is displayed.
2. Enter a name for the mapping set
3. The *Define Mapping Set* dialog box is divided into three tabs:
  - *Calculated Fields*  
Use this tab to map calculated fields to metadata in the field source.
  - *Basic Metadata*  
Use this tab to map standard metadata (such as *Last Modified Date*) to metadata in the field source.
  - *Standard Mappings*  
This tab explains how the basic file properties: *Created date*, *Last modified date*, *Last accessed date*, *File owner*, *File name*,



are automatically mapped by Discovery Center following file migration. The values are set according to the strategy outlined in the following table.

Table 16 Mapping of basic file properties following migration

Property	With File System Connector	With SharePoint Connector	Default Behaviour
<i>Created date</i>	Set to the value from the source file at time of migration	Set to the value from the source file at time of migration	Set to the date and time that the file was moved
<i>Last modified date</i>	Set to the value from the source file at time of migration	Set to the value from the source file at time of migration	Set to the date and time that the file was moved
<i>Last accessed date</i>	Set to the value from the source file at time of migration	N/A	Set to the date and time that the file was moved
<i>File Owner</i>	Set to the account credentials used for the move action.	Set to the account credentials used for the move action.	Set to the account credentials used for the move action.
<i>File name</i>	Unchanged unless modified to avoid duplicate file names	Unchanged unless modified to avoid duplicate file names	Unchanged unless modified to avoid duplicate file names

4. Select the *Calculated Fields* or *Basic Metadata* tab, as required.
5. From the *Source* dropdown list, select the Discovery Center metadata item to be mapped to a field in the repository.
6. From the *Destination* dropdown list, select the repository's corresponding metadata field.

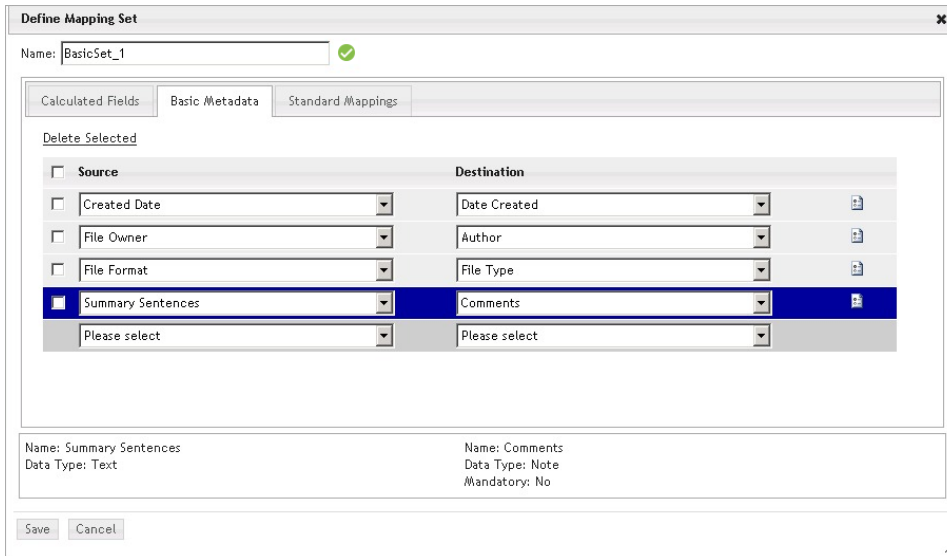



Figure 94 Basic metadata mappings

7. By default, Discovery Center does not migrate a file if the destination field is missing. If a file does not have a value for any metadata fields in the mapping set, then the Discovery Center will attempt to migrate the file without setting a value. If the unset field is a mandatory requirement, then an error will be recorded and the file will not be migrated.
8. To choose alternative actions in these circumstances, click on the  icon on the right. The *Mapping Options* dialog box is displayed.



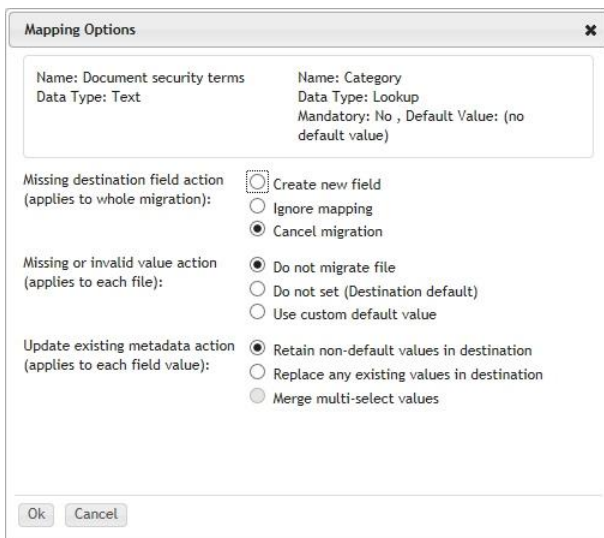


Figure 95 Mapping Options

9. Choose the required action in the event that a file does not have a value for the mapped metadata, or the value is not valid for the destination. This rule is applied on a file-by-file basis:
    - **Create new field**  
Migrate the file, creating the necessary metadata field in the destination file store.
    - **Ignore mapping**  
Migrate the file. The metadata mapping is lost.
    - **Cancel migration**  
(Default) The file is not migrated.
  10. Choose the required action in the event that the file does not have a value for the mapped metadata:
    - **Do not migrate file**  
The file is not migrated.
    - **Do not set (Destination default)**  
(Default) Migrate the file and apply the default field value as set in the destination repository.
    - **Use custom default value**  
Migrate the file and apply a custom value to the destination field. When you select this option, a text box is displayed for you to enter the custom value.
  11. Choose the required action for updating metadata (the rule is applied on a field-by-field basis):
    - **Retain non-default values in destination**  
(Default) Replace the existing value if it is the default but retain any custom value.
    - **Replace any existing values in the destination**  
Replace the existing field whether it is a default or custom value.
    - **Merge multi-select values**  
Combine the new and existing entries.
  12. Click on the **OK** button to save the mapping options and return to the mapping set.
  13. Repeat this procedure for all required metadata mappings on the *Calculated Fields* and *Basic Metadata* tabs. Click on the **Save** button when you have finished.
- To delete a mapping, select its check box and then click on the **Delete selected** link at the top of the page.

**Note.** You can map Discovery Center metadata to two or more fields in the repository but it is not possible to map multiple Discovery Center metadata to the same repository field.





To edit an existing mapping set, click on its edit icon in the *Actions* column.

#### Deleting a Mapping Set

To delete a mapping set, select its check box and then click on the Delete selected link at the top of the page.

#### Supported Metadata Field Types

*Appendix 3: Preparing for SharePoint Migration* on page 182 provides details of the SharePoint field types supported for migration mapping (and discovery).

## Character Mappings

When an Information Manager chooses to migrate files as a reporting action, the use of a Character Mapping Set ensures that illegal characters are removed or replaced with acceptable text or other characters.

For example, the following characters are not permitted in SharePoint Sites or Document Libraries:

/ \ : \* ? " ' < > | # <TAB> { } % ~ &

(See [support.microsoft.com/kb/905231](http://support.microsoft.com/kb/905231) for full details).

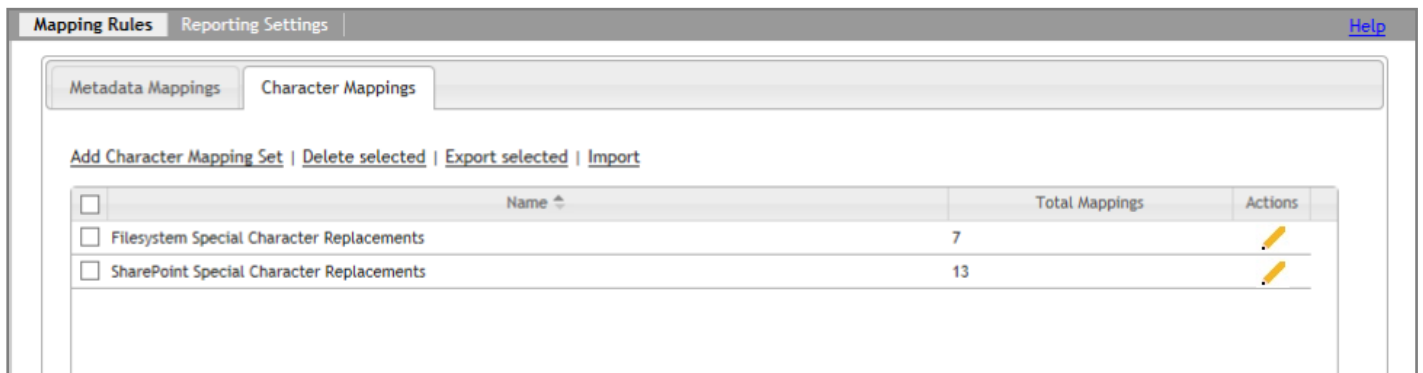


Figure 96 Reporting and Actions page – Mapping Rules – Character Mappings



Discovery Center is supplied with two character mapping sets as described below.

Table 17 Standard Character Mappings Sets

Filesystem Special Character Replacements	SharePoint Special Character Replacements
:	~
*	#
?	%
“ ‘	&
>	*
<	} )
	{ (
	:
	>
	<
	?
	” ‘



## Adding a Character Mapping Set

1. Click on the **Add Character Mapping Set** link. The *Define Character Mapping Set* dialog box is displayed (see **Error! Reference source not found.**).
2. Enter a Name for the Mapping Set.
3. Click on the Add Mapping link. Use the Selected Mapping box to define a mapping action:
  - **Replace in**  
Choose to carry out this action on:
    - File names
    - Folder names
    - File and Folder names
  - **Target characters**  
Enter the character or character string to be removed or replaced during migration.
  - **Mapping action**  
Choose from the following actions:
    - **Remove**  
Extract the character without any substitution.
    - **Replace**  
Substitute the character with an alternative character or string.
    - **Replace with**  
(displayed for Replace mapping action only) Type the alternative character or string to replace the target character(s).
4. Click on the **Save** button. The mapping action is listed in the *Character Mappings* box.
5. Repeat this procedure to add further mapping actions to the Character Mapping Set.
6. When you have defined all the mappings for this set, click on the **Save** button in the lower left corner of the dialog box. The Character Mapping Set is now listed on the *Character Mappings* tab.

**Note.** Replacing or removing characters can have unexpected effects. For example:  
The removal of a character, such as a space, may result in the replication of an existing file name or folder name.  
The replacement of single characters (such as spaces) with longer character strings (for example, "%20") could result in file, folder or path names exceeding size limits.  
Ensure that you fully understand the consequences of applying a Character Mapping Set before using it.

## Editing Mapping Actions

To edit an existing mapping action, select it in the *Character Mappings* list and then click on the **Edit** link in the *Mapping Details* box. Change the mapping as required and then click on the **Apply** button.

## Deleting Mapping Actions

Select the check boxes of the mapping sets to be deleted. Then, click on the **Delete selected** link.

## Editing a Character Mapping Set

To edit an existing mapping set, click on its **Edit** icon on the *Character Mappings* tab. The *Define Character Mapping Set* dialog box is displayed. Follow the procedures listed above to add, edit or delete the set's mapping actions.



## Deleting Character Mapping Sets

Select the check boxes of the mapping sets to be deleted. Then, click on the **Delete selected** link at the top of the page.

Define Character Mapping Set ✕

Name:  ✓

### Character Mappings

[Add mapping](#) | [Delete selected](#)

<input type="checkbox"/>	Target	Action	Where
<input checked="" type="checkbox"/>	~	Replace with '_'	File and Folder Names
<input type="checkbox"/>	#	Replace with '_'	File and Folder Names
<input type="checkbox"/>	%	Replace with '_'	File and Folder Names
<input type="checkbox"/>	&	Replace with '_'	File and Folder Names
<input type="checkbox"/>	*	Replace with '_'	File and Folder Names
<input type="checkbox"/>	}	Replace with ')	File and Folder Names
<input type="checkbox"/>	{	Replace with '('	File and Folder Names
<input type="checkbox"/>	:	Replace with '_'	File and Folder Names
<input type="checkbox"/>	>	Replace with '_'	File and Folder Names
<input type="checkbox"/>	<	Replace with '_'	File and Folder Names
<input type="checkbox"/>	?	Replace with '_'	File and Folder Names

### Mapping Details

Replace in:  
File and Folder Names

Target characters:  
~

Mapping action:  
Replace with '\_'

Replace with:  
-

[Edit](#)

[Save](#) [Cancel](#)

Figure 97 The SharePoint Special Character Replacements mapping set



# Reporting Settings

### Reporting Database Processing

The process reporting database task is currently active.

[Edit](#) [Process Now](#)

### Management Reporting Database Processing

Automatic processing is enabled. A task to process the database will be scheduled at 1700 hours on Monday after any indexing, action, file metadata import or location deletion task is completed.

Last processed: 2018/06/04 09:35  
Times displayed in (UTC) Coordinated Universal Time

Instance name: localhost  
Database name: ActiveNavigationManagementReportData  
Credential Type: Service Account(s)  
Credential: Use the account specified for the Scheduler and the Web Application

[Edit](#) [Process Now](#)

### Reporting Settings

File list limit	1,000,000	<a href="#">i</a>
File list cache limit	20	<a href="#">i</a>
Enable Quick Load mode for Saved Views	true	<a href="#">i</a>
Action error threshold (percentage of total files)	10	<a href="#">i</a>
Export File Type	Xlsx	<a href="#">i</a>

[Edit](#)

### Quarantine Locations

[Add Quarantine Location](#) | [Delete Selected](#)

<input type="checkbox"/>	Name ↕	Quarantine Location
<input type="checkbox"/>		\\review\test new docs\localhost\test old docs\

Figure 98 Reporting and Actions page – Reporting Settings



## Reporting Database Processing

The Discovery Center builds reports from a reporting database stored and managed using SQL Server Analysis Services. The reporting database is independent of the indexing database and is optimized for reporting performance. As a consequence, the reporting database needs to be updated as indexes change, either as a result of new index tasks, actions from reports or metadata import tasks.

Processing the reporting database synchronizes it with the contents of the indexing database, and this may take some time depending on the amount of data that has been indexed. Leave automatic processing enabled unless the time taken to process the reporting database is long and you need a sequence of actions to complete before you intend to report from the system.

**Note.** It is important to note that all reporting functions will still succeed when the reporting database is out of date. For this reason, there is no need to process the reporting database after every activity. The default configuration will automatically add a task to process the reporting database after groups of index processing, metadata import, or action tasks have been completed.

Click on **Edit** to change the settings controlling the generation of the reporting database processing task. Choose from the following three options:

- **Automatically** (default)  
After any indexing, action or file metadata import activity is completed, a task to process the reporting database will be added at the end of the Current Activity queue (if it is not in the queue already).
  - **Prioritization delay (hours)**  
Set the maximum number of hours (default: 12 hours) after any indexing, action or file metadata import activity is completed until the Processing Reporting Database task is moved or added to the top of the queue. Before this time, the task will be added as the last item in the Current Activity queue (if not present already) after an activity is completed.
- **Manually**  
Stops the scheduler from automatically queueing or scheduling the Process Reporting Database task after the completion of any indexing, action or file metadata import activity. This can save time when you expect to add many long-running items in to the Current Activity queue and do not need to view reports until all of the work is finished. You will need to process the reporting database manually by clicking on the **Process Now** button. Re-enable automatic processing if required.
- **Scheduled Daily**  
Process the Reporting Database at the same time each day.
  - **Schedule Reporting Database Processing Daily at:**  
Choose the hour of the day (default: 00:00) when the Processing Reporting Database task is added to the Current Activity queue.

You can reposition the task to process the reporting database in the queue: if it has not yet been automatically added then click the **Process Now** button to add the task, and go to the *Current Activity* tab to move it to the required position.



## Management Reporting Database Processing

The Management Reporting Database is only available if your license solution includes the Management Reporting Feature Pack (see page 189). The database is not created automatically and must be set up separately using a batch file supplied during Discovery Center installation.

1. Click on **Edit** to change the settings controlling the generation of the Management Reporting database processing task.
2. Set the **Instance name** and **Database name**. These settings are checked and will only be accepted if connection can be established and if the database has the expected structure.
3. Select a **Credential Type**. This setting determines whether the connections made to the Management Reporting Database are made using Windows or SQL authentication.
  - **Service Account(s)** (default)  
Connections to the Management Reporting Database are made with Integrated Windows authentication using the account(s) specified for the Scheduler and the Web Application.
  - **SQL**  
Connections to the Management Reporting Database are made using the specified SQL authentication:
    - **Credential**  
Choose from the credentials displayed in the drop-down list (as defined in *System Settings > Credential Management*, see page 49).
4. Choose the schedule for processing the Management Reporting database:
  - **Automatically** (default)  
The database is updated according to the specified schedule. By default, the database is scheduled for automatic weekly updates at 2am each Sunday.
  - **Manually**  
Stops the scheduler from automatically queueing or scheduling the Management Reporting Database Processing task. You will need to process the database manually by clicking on the **Process Now** button. Re-enable automatic processing if required.
  - **Schedule Management Reporting Database Processing Daily at**  
Choose the hour of the day (default: 02:00, Sunday) when the Management Reporting Database Processing task is due to be added to the Current Activity queue. This will only occur after the completion of any indexing, action, file metadata import or location deletion task. Only available when automatic processing has been selected.

You can reposition the task to process the Management Reporting Database in the queue: if it has not yet been automatically added then click the **Process Now** button to add the task, and go to the *Current Activity* tab to move it to the required position.

**Note.** For information about creating reports from Management Reporting data, please refer to page 189.



## Reporting Settings

This section contains the following options:

- **File list limit**  
In *Report Viewer*, the *File List* tab (see page 146 for more details) displays paged information and metadata for all the files in the current report. Use this setting to specify the maximum number of files to be listed (default: 1,000,000, maximum value: 9,999,999).
- **File list cache limit**  
To minimize the time spent collating File Lists for reports, and the time spent moving between pages in the file lists, the most recently viewed lists are cached in the database. The maximum number of lists that will be cached is set here, and can be increased or decreased if necessary (default: 20, maximum value: 1000). To cache a list of 1,000,000 files will require approximately 24MB of space in the database.
- **Enable Quick Load mode for Saved Views**  
Select this option to cache saved views following processing of the Reporting Database.
- **Action error threshold (percentage of total files)**  
Use this setting to set a limit to the number of errors allowed during an action before it is aborted. The threshold is calculated as a percentage (10% by default) of the total number of files actioned. Use a value of 100 to force an action to complete irrespective of the number of errors. The exact number of errors permitted before an action aborts depends on the batching of actions and the number of threads.
- **Export File Type**  
Choose the default file format for the export of file lists, data tables, audit trails, activity histories and metadata value files: CSV (default) or XLSX.

Click on **Edit** if you want to change these settings.

## Quarantine Locations

Specify one or more folders to be used as quarantine locations for files deleted as a result of a reporting action in Discovery Center (see page 155).

**Note.** The quarantine location will require careful management to ensure that files can be restored, if necessary, and all genuine ROT is discarded after a suitable time period. If the quarantine location is included in an index, its contents will contribute to the advisory *Volume Under Management* total.





# Appendix 1: Indexing and Analysis

## Overview

When managing Discovery Center indexes, it is important to understand the work that the Discovery Engine is undertaking so that the time and resources needed to complete any given task can be understood and planned. Further, when a problem is encountered, a clear understanding of how all indexing processes and tasks work together will be crucial to successful diagnosis and corrective action.

When an index is run, the containers and files within that index are processed according to the index configuration. Indexing always proceeds in the order below although steps may be omitted according to index settings as shown.

Index Process	Tasks	Index Configuration	
Skim	Gather file topography and basic metadata	Always re-skim	Always occurs if this index option is checked
Analysis	Analyze files for ID, duplicates and contents	Skim, Duplicate and Textual	Choose index options
Field Calculation and Classification	Update field values from extraction rules values; Update classifications from index results		
Report Database Processing	Prepare database for reporting	Process Reporting Database	Check the process reporting database index option

Figure 99 Index processing steps



## Skim Processes

Skimming always gathers file and basic metadata, including file names, file path, sizes, dates and extensions. Optionally, skimming may also collect file owner and repository property information. Note that a successful skim is required before any analysis can proceed.

## Analysis Processes

All analysis processes take place according to index configuration. Some processes require that others have been completed before they can proceed. The flow chart below shows how analysis processes occur relative to each other and their specific dependencies.

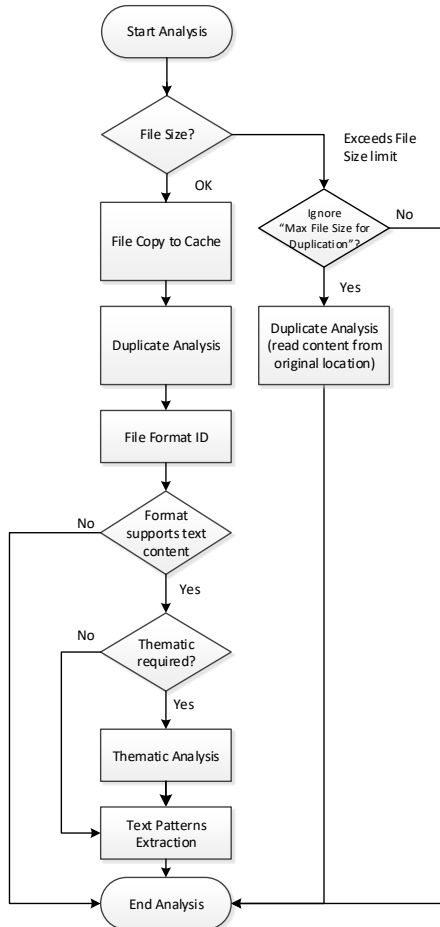


Figure 100 Analysis processing flow chart

### File Retrieval

Before analysis can take place, copied of files must be retrieved from storage and written to the Discovery Center File Cache at *install location/Discovery Center/File Cache*. The process will take place at the maximum speed permitted by the network environment and the Discovery Center host server disk input/output performance. Note that file cache contents are removed after each file has been analyzed.



## Duplicate Analysis

The cached copy of each file is analyzed by a SHA256 hashing algorithm. The resulting hash value is stored in the index results.

## File Identification

File identification determined the exact version of the application which last saved the file to be analyzed. The type of files which can be identified are determined by the Oracle OutsideIn libraries included with the Discovery Center. For more information see [https://docs.oracle.com/en/middleware/standalone/outsidein/8.5/oit-supported-fileformats/OutsideIn\\_8.5.5\\_Supported\\_File\\_Formats.pdf](https://docs.oracle.com/en/middleware/standalone/outsidein/8.5/oit-supported-fileformats/OutsideIn_8.5.5_Supported_File_Formats.pdf).

Successful file identification is required before a file can be converted for content analysis. File contents analysis cannot be performed on files that cannot be converted.

## File Contents Analysis

**File Conversion.** Before a file can be analyzed it must be converted to text or HTML by the Oracle OutsideIn libraries. Any failure in file conversion usually indicates that the file is corrupt, encrypted, password protected, or its file type is not supported. Failures to convert files are recorded in the Discovery Center analysis logs.

**Thematic Analysis.** Text in converted files is analyzed for its thematic contents; thematic analysis also produces file summaries and the raw results required to identify textually similar files. Thematic analysis can handle text in English, French, German, Italian or Spanish.

**Text Pattern Extraction.** Text pattern and keyword extraction is performed using content regular expressions defined in metadata extraction rules and applied to a calculated field. Note that the use of many content regular expressions can have a significant impact on analysis performance; the time taken for this work is recorded at the end of each index analysis in the analysis logs.

## Other Analysis Types

**File Path Text Patterns.** Regular expressions are defined in metadata extraction rules to extract text patterns for file paths. These rules run very quickly.

**Embedded File Property Extraction.** Values are extracted from embedded Windows or EXIF file properties as defined in the corresponding metadata extraction rule.

**Repository Property Extraction.** Named properties for files in connected repositories (such as Microsoft SharePoint) are configured using extraction rules applied to calculated fields. The performance of this activity depends upon the performance of connections to the relevant repository.



## Index Post-Processing

Index post processing includes a number of tasks required to collate the results of skimming and analysis and make those results available for reporting and action.

**Classification.** Where indexes contain calculated fields requiring classification, those values will be calculated initially after skim and then subsequently after analysis has been completed.

**Reporting Database Processing.** Index results are stored in the ActiveNav SQL transactional database, optimized for data collection and actions. For reporting, however, the Discovery Center transforms index results into SQL Server Analysis Services database. By default, a task to process the reporting database will be placed on the current activity queue after any index processing or action task has completed, with a priority controlled by configuration in the *Reporting Settings* tab of the *Reporting and Actions* page (see page 173). This ensures that reports show up to date information.



# Appendix 2: Metadata Model

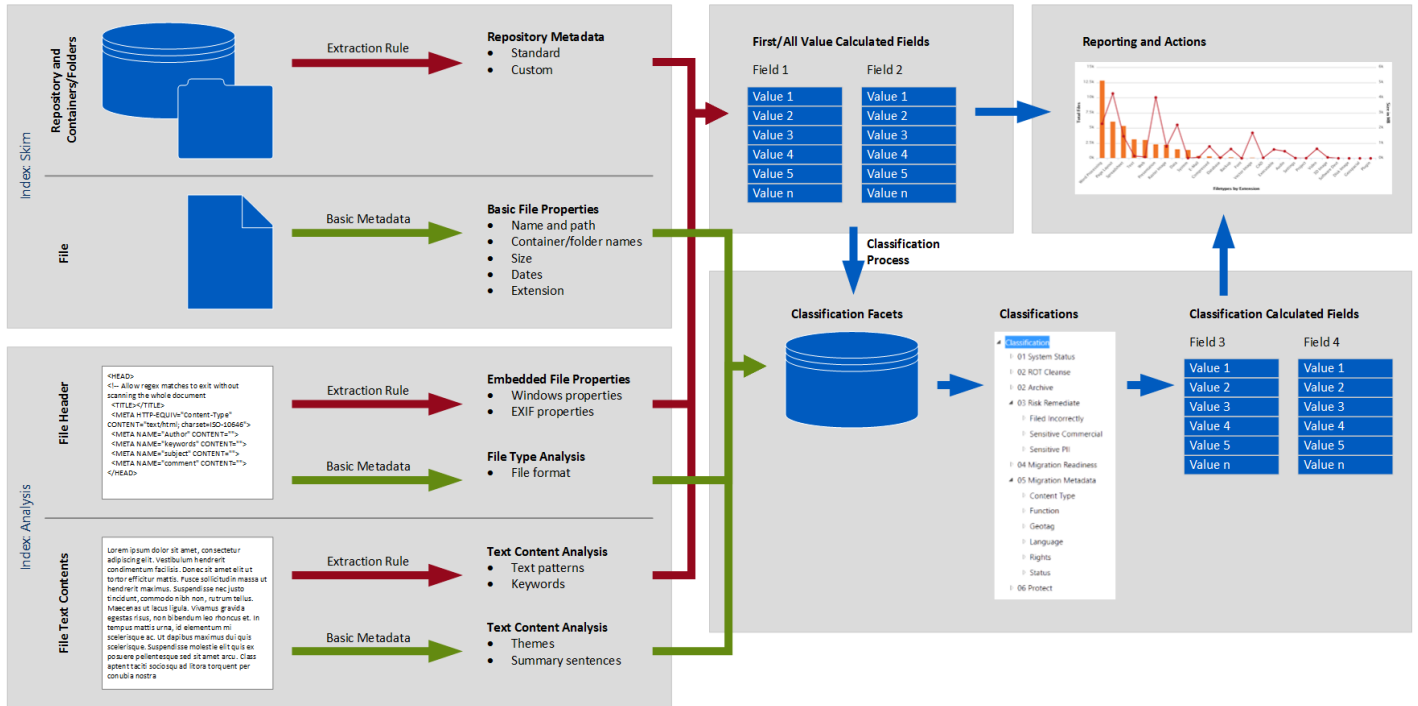


Figure 101 Metadata model



# Appendix 3: Preparing for SharePoint Migration

To get the best results from the migration process you should:

- Carefully consider the requirements for metadata.
- Prepare your content to ease the migration process.
- Build a structure that will meet the specific restrictions of SharePoint.
- Filter out content that cannot be uploaded.

## Preparing SharePoint

To prepare your SharePoint site:

1. **Decide on the structure.**  
In advance of migration, identify the library structure you want to achieve so that you can then apply this structure when organizing content.
2. **Plan the metadata.**  
Once you have determined the relevant Document Libraries and associated structure, you can then assess the metadata that should be used in these libraries. When planning metadata for bulk migration of legacy content, you must determine an acceptable balance between the metadata that can be extracted from your content and the needs of the business.
3. **Define content types.**  
Decide if you would benefit from defining a range of content types for your libraries.
4. **Create relevant columns.**  
When you have determined the required structure and metadata attributes, we recommend that you create the necessary Metadata columns at SharePoint Site Collection level. This allows the metadata values to be re-used throughout the SharePoint Environment. If the metadata is only set up within a document library, the metadata values will only exist at the document library level and will not exist at a site level.
5. **Review SharePoint defaults.**  
SharePoint imposes limits on acceptable file types and sizes; you may want to review the default limits to determine if they are suitable for your needs.



## Preparing Content

The results of SharePoint planning can be used to guide the use of the ActiveNav tool set to prepare content for migration.

### 1. **Work towards the metadata requirements of the SharePoint site.**

The metadata requirements for the destination SharePoint site(s) must be used to guide the configuration of Discovery Center's metadata analysis process. This must be finalized before the Discovery Center index of the data can be created.

**Note.** If the SharePoint site will contain mandatory metadata columns then you must prepare relevant attributes to assist in meeting these requirements.

### 2. **Prepare content to fit SharePoint's restrictions.**

There are a number of restrictions that may affect the success of a migration

- Document size
- Permitted file extensions
- Length of folder names/filenames/complete path length:
  - 256 Unicode (UTF-16) characters - the effective file path length limitation, including a domain/server name
  - 128 Unicode (UTF-16) characters - the path component length limitation

The following links provide a good introduction to these issues. Some restrictions such as permitted file types and file sizes can be adjusted with care.

<http://technet.microsoft.com/en-us/library/ff919564.aspx>

<http://support.microsoft.com/default.aspx?scid=kb;en-us;905231>

### 3. **Remove invalid characters from file, folder and site names.**

Carry out a find and replace session to identify and remove any illegal characters: # % & \* : [ ] ? / \ { | } ~. For example, some organizations use the ampersand (&) extensively in filenames. This will cause problems when migrating data into SharePoint. It is best to tackle these issues up front by carrying out a find and replace session, replacing invalid characters and then crawling the content to extract the metadata.

You can also create character mapping sets to remove or substitute illegal characters during migration (see page 169).

### 4. **Organize content to reflect the desired structure within SharePoint.**

If possible, align content with the document library structure so that content can be migrated to multiple document libraries in one go.

### 5. **If you have put the effort into planning, then migration should be a very straightforward process.**

6. If you are migrating a significant amount of content, or if your metadata requirements are complex, then it is possible that some content will not be migrated successfully at the first attempt. Also, be aware that files located in a document vault can cause a migration failure owing to access delays.



## Field Types Supported for Migration

Table 18 SharePoint Field Types supported for migration

Field Type	Description
<b>Content Type</b>	Files moved to SharePoint will have their content type set according to metadata mapping rules defined by the user. On migration the SharePoint connector will seek a suitable content type by matching its name with the value from the mapped calculated field. A value will be written from the mapped calculated field as long as the field value is valid according to the constraints defined in SharePoint for the destination library.
<b>Managed Metadata Term Set (single/multi)</b>	Calculated field values will be written where they match the column's managed metadata term set hierarchy. If the calculated field contains multiple values for a file, all values will be written according to the SharePoint field setting.
<b>(Same Site) Lookup (single/multi)</b>	Each value from the mapped calculated field will be written if it matches an available lookup source value. Note that lookups to different sites are not supported.
<b>Choice (single/multi)</b>	Each value from the mapped calculated field will be written if it matches a valid choice. If the Allow fill in choices is set in the target field, new values will be written into the field choices.
<b>Date/Time, Number, Currency</b>	Discovery Center will attempt to assign the calculated field value to the mapped field; write will fail if the value is not valid for the target field type.
<b>Yes/No (Boolean)</b>	Yes/No values in the source calculated field will be written to the mapped field.
<b>Hyperlink</b>	A string will be written from the mapped calculated field.
<b>Text (single line of text)</b>	A string will be written from the mapped calculated field. The string will be concatenated according to the SharePoint field configuration, and must be within the configured maximum length.
<b>Text (multiple lines of text) also known as Note columns</b>	A string will be written from the mapped calculated field. By default, multi-line columns have a 255 character limit unless the "Allow unlimited length in document libraries" option has been chosen in the document library.

## Field Types Not Supported for Migration

Table 19 SharePoint Field Types not supported for migration

Field Type	Description
<b>User</b>	User fields can be mapped to but a successful user mapping requires a text match between the source calculated field and the target Active Directory lists. This function is not supported but may be achievable with suitable source data.
<b>External data</b>	It is not applicable to write to an external data field type as its values depend upon the connected external data source.
<b>Managed metadata (enterprise keywords)</b>	The SharePoint Web Services API used by the SharePoint connector does not support writing enterprise keywords.
<b>Calculated</b>	It is not applicable to write to a calculated field type as its values depend upon the values of other fields.





# Appendix 4: Files Supported for MIP Sensitivity Labeling

The files that can be labeled through Discovery Center are driven by restrictions enforced in the MIP integration components provided by Microsoft and by restrictions stated in Microsoft documentation. These restrictions are detailed in this appendix.

In all cases where a label has not been applied due to these restrictions an appropriate message will be presented to the user in the [File Labeling Status Report and Audit Trail](#) details for the report action.

## File Format & Extension Restrictions

Restrictions on File Formats and File Extensions are enforced both by Discovery Center and by the MIP integration components. At the time of writing, Microsoft documentation on supported file types for MIP can be found at: <https://docs.microsoft.com/en-us/information-protection/develop/concept-supported-filetypes>

This documentation separates support for labeling different types of file into three categories:

1. File types supported for labeling;
2. File types which only support labeling when the label applies RMS Encryption. In these cases, the extension of the labeled file will be updated to protected file extension denoted by a 'p' prefix (for example `.txt` becomes `.ptxt`)
3. File types that can never be labeled.

Further to this, restrictions on applying labels through Discovery Center may be encountered when:

- The file extension is permitted, but the format of the file is inconsistent with that extension;
- The file format is permitted, but the file extension is inconsistent with this format;
- No extension is applied to the file.

## Other Labeling Restrictions

Applying a MIP Sensitivity Label through Discovery Center may also be restricted for the following reasons:

- The file is empty (i.e. has a size of 0B);
- The label to be applied matches the label already applied to that file;
- The label to be applied would downgrade the sensitivity of the document, but the action performing the label has not been configured to allow downgrade;
- The value supplied for labeling does not correspond to a label in a Label Policy accessible to the credential supplied for labeling, as documented in the [MIP Settings](#) section.
- The label you have attempted to apply has been made Inactive in the Label Policy;



# Appendix 5: Optimizing Configuration of Key Microsoft Components

The following describes optional steps that may be taken to improve Discovery Center performance by optimizing the configuration of key components such as SQL Server or IIS. Note these instructions involve detailed administration of those components and should be executed in consultation with local systems administrators to ensure policies are not violated.

## Enable IIS Dynamic Data Compression to Reduce Impact of Data Export on Network

It is possible to configure IIS to use data compression between the server and web browser. This decreases the amount of time taken to transmit data across a network at the expense of slightly increased CPU load on the server and client machines. It is of most benefit across a slower network connection. Compression is not enabled for dynamically generated content (like the Discovery Engine pages) in a default IIS installation.

Network bandwidth of 20 Mbps has been measured during file export from a moderate SQL server installation. If this is likely to exceed the network capacity between your Discovery Engine and the users downloading exported data then you should enable dynamic content compression. The steps to do this for IIS 7 using the IIS Manager application are shown below:

- Use Server Role Manager to install Static Content Compression and Dynamic Content Compression. In Server Manager, navigate to Server Manager, Web Server (IIS) and click Add Role Service
- If they are not both already selected select Static Content Compression and Dynamic Content Compression in the list of role services, and install them by clicking Next.

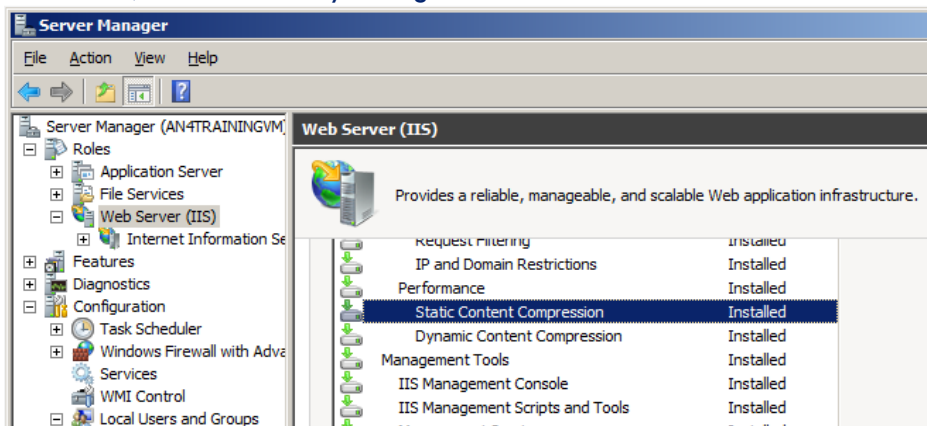


Figure 102 Installing Static Content Compression

- Use the IIS Manager application to turn on and configure the compression options for the server, as shown below.



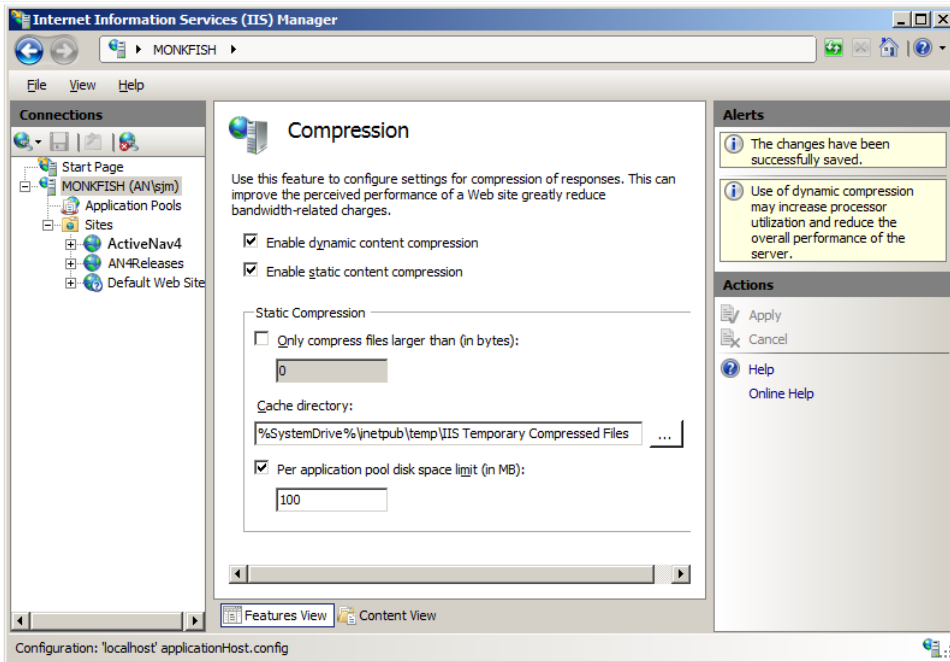
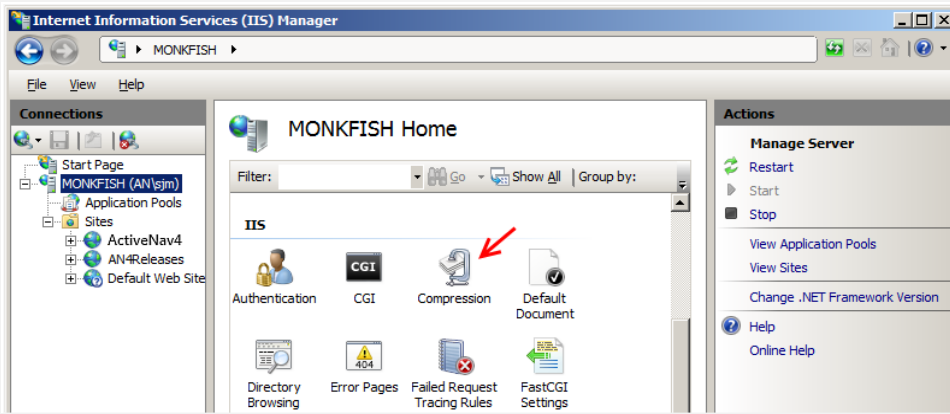


Figure 103 Configuring Compression options

- Ensure that dynamic and static compression are also enabled for the Discovery Center web site (ActiveNav4 by default).

## Configure SQL Server for optimal performance

See the knowledge base article “Best Practices for SQL Server Configuration” for recommendations on SQL Server configuration. These recommendations combine general best practice advice and specific lessons learned from Discovery Center deployments.



## Configure Database Volume Maintenance Rights

Windows Server provides facilities to allow trusted accounts to allocate large files in disk instantly. For Discovery Center, this has specific relevance to improve the way SQL Server allocates and manages space for the databases used by Discovery Center (including the tempdb and associated transaction log files). It is therefore considered good practice to configure the SQL Server database service account with the rights required to perform volume maintenance. To set up volume maintenance rights:

- Determine the identity of the SQL Server Service Account.
- Locate the **Perform volume maintenance tasks** policy in the Local Security Policy Manager.
- Add the SQL Server Service Account to the **Perform volume maintenance tasks** policy.

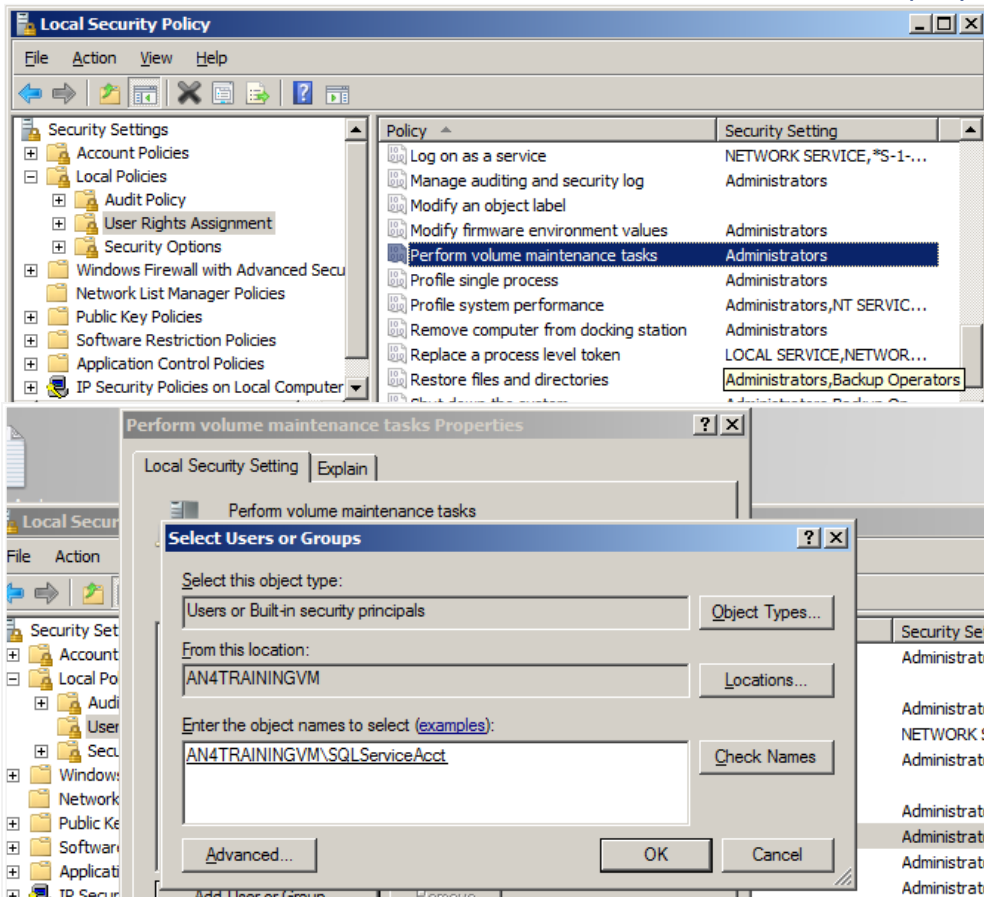


Figure 104 Configuring database volume management rights

# Appendix 6: Preserving NTFS File Owner

When performing an MIP Sensitivity Label action against a Windows File System the NTFS File Owner value for the labelled file may be set to the credential being used to perform the action.

Discovery Center contains functionality to detect these changes to File Owner and attempt to revert the File Owner to its previous value.

This functionality can be enabled or disabled by updating the **MIPLabelActionPreserveFileOwner** and **MIPLabelActionErrorOnFileOwnerUpdateFailure** configuration values in the **Scheduler.config** file.

The functionality is enabled by default, however it may be beneficial to disable it for the following reasons:

You know that the credential being used for the MIP Sensitivity Label action, or the environment where the action being performed does not meet the criteria for File Owners to be updated. For more information on these requirements, refer to **Appendix 12** of the **Discovery Center Installation Guide**.

If preserving File Owner is not required for your labelling scenario, disabling this functionality will increase the speed of the MIP Sensitivity Label action.

## MIPLabelActionPreserveFileOwner

If this setting contains a value of **True**, then following an MIP Sensitivity Label action, Discovery Center will detect if the file owner has changed following labelling and attempt to update the File Owner.

In the event that labelling has updated the file extension, resulting in a new file, Discovery Center will attempt to apply the owner value to the new file.

If this setting contains a value of **False**, then Discovery Center will take no action to ensure the File Owner is unchanged following labelling.

The default value for this setting in a new installation or upgrade of Discovery Center is **True**.

## MIPLabelActionErrorOnFileOwnerUpdateFailure

If this setting contains a value of **True**, then in the event Discovery Center attempts to preserve File Owner information following labelling, but is unable to do so, an error will be recorded in the Process Statistics of the MIP Sensitivity Label report action.

If this setting contains a value of **False**, then no error will be recorded in Process Statistics (but will still be present in the Scheduler log output).

The default value for this setting in a new installation or upgrade of Discovery Center is **True**.



# Appendix 7: Management Reporting

Management reporting is intended to provide project and information governance teams with the means to automatically aggregate results from their Discovery Center instances over time so that they can create reports and dashboards to show, for example:

- Overall storage metrics including cost, risk and other policies
- Historical data on the metrics over time.
- Actions on content taken over time through Discovery Center.

These features are designed to be used with a customer's chosen third party business intelligence and reporting products.

## Overview

Management reporting is activated like other optional Discovery Center features by applying the relevant license file.

The Discovery Center application ships with a script that enables the deployment of a Management Reporting Database (MRD) to SQL Server, independent of the rest of the application. Each independent Discovery Center installation can be configured to reference an existing MRD so that it can connect and upload its data to a shared instance if required.

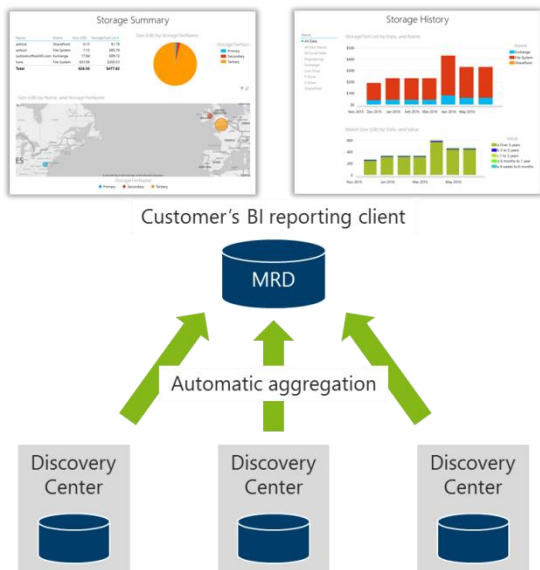


Figure 105 Management Reporting Database Architecture

After installation is complete, each connected Discovery Center can automatically pass its data to the MRD on a scheduled weekly basis, or when manually initiated.

Once data is available in the MRD, a compatible reporting tool can be connected and synchronized according to its vendor's instructions.



# MRD Design

The MRD is comprised of three types of table.

- Import tables support the database synchronization process and should not be used for reporting.
- Dimension tables provide key reporting dimensions covering:
  - Dates for each time slice of data.
  - Discovery Center instances.
  - Servers for repositories.
  - Indexes.
  - Areas of interest.
  - Network map locations related to fact table data.
  - Connectors.
  - Storage tiers.
  - Calculated field names.
  - Report types.
  - Action types.
  - Report selections.
- Fact data tables covering:
  - Data volumes, costs, coverage statistics and counts for calculated fields by index and area of interest.
  - Data volumes, costs, and counts for system activities such as report actions and index processes.

## Discovery Center Configuration

### Calculated Fields

Management reporting automatically collects file volume and count data for indexes, areas of interest and activities. In addition, selected calculated can be included in management reporting as follows:

1. Go to Discovery Center > Metadata > Calculated Fields.
2. For each field you wish to add to management reporting check **Available for management reporting**.

### Storage Tiers

To include storage costs in management reports, ensure that storage tiers have been configured for each Discovery Center:

1. Go to Discovery Center > Network Map > Storage Tiers.
2. Add new tiers and assign storage cost metrics.
3. Go to **Discovery Center > Network Map**. Edit locations in the network map hierarchy to apply tiers in turn.

### Repository Server Geographic Location

Geographic locations for repository servers can be used to support map based presentation of data. Geographic locations must be added direct to the underlying management reporting data model. Using SQL Server Management Studio or similar, locate the Servers table and add latitude/longitude values for each server in the appropriate fields. This must be done after the Management Reporting Database has been processed to include the servers you wish to configure.



# Reporting Tool Data Configuration

Detailed steps for reporting tool configuration depend upon vendor directions. The following provides example steps for Microsoft PowerPivot.

## PowerPivot Essential Set Up

### Connect

1. Enable PowerPivot in Excel.
2. Open PowerPivot.
3. Choose Get External Data > From Database > From SQL Server.
4. Connect to your Management Reporting Database.
5. Select fact and dimension tables for import and complete the import.

### Set Up Date Dimension

1. In PowerPivot select the DateDimension table.
2. Click Design > Mark As Date Table > Mark As Date Table.
3. Select the *Date* option and click **OK**.

## PowerPivot Optional Set Up

### Configure Fact Units

For each facts table:

1. Add a column to convert bytes to GB (or TB as needed).
2. Set the data type for Storage Tier currency.

### Add Labels To DateDimension Table

1. Add a column named *QuarterName*. Set the column calculation to `'Q' & [Quarter]`
2. Add a column named *MonthName*. Set the column calculation to `'SWITCH([Month], 1, "Jan", 2, "Feb", 3, "Mar", 4, "Apr", 5, "May", 6, "Jun", 7, "Jul", 8, "Aug", 9, "Sep", 10, "Oct", 11, "Nov", 12, "Dec", "Unknown month number")'`
3. Add a column named *DayName*. Set the column calculation to `'SWITCH ([DayInWeek], 1, "Sun", 2, "Mon", 3, "Tue", 4, "Wed", 5, "Thu", 6, "Fri", 7, "Sat", "Unknown day")'`

# Starting Report Design

## Reporting on Data State

When designing a report, decide first on the scope of the report – that is, what to ‘report on’. The MRD contains two different scopes of date – the first includes all files indexed, the second includes only those files which have a value for a calculated field set as ‘Available for management reporting’ (see above).

Once scope is understood, choose whether you wish to report with reference to indexes and servers or areas of interest. The matrix below shows which facts table to use according to your chosen scope and reference.





With the correct facts table selected, reports can be created with reference to any linked dimension table as shown in the schemas below.

Table 20 Scope of Facts tables

		Report On	
		All Files Volume and Count	Only Files with Calculated Field Values
Report Reference	Indexes and Servers	facts.VUMIndexes	facts.CalculatedFieldIndexes
	Areas of Interest	facts.VUMAreaofInterests	facts.CalculatedFieldAreaofInterests

### Reporting on Activity History

The MRD contains data about activities taken in Discovery Centers which affect the number/volume of files under management. There is only one facts table available for this purpose; that table shares linked dimensions with the above facts tables, as shown in the schemas below.

### Working with Dimensions

Tables in the dim.\* schema are related to the facts.\* tables so that you are able to slice, filter and aggregate data by key properties in the way most relevant to your business reporting needs.

Normally, your Business Intelligence tool will be able to utilize these tables and their relationships to help you build the relevant dashboards to highlight the findings of your Discovery Center implementation and in some instances will enable you to build interactive reports that allow users to explore the data held in the Management Reporting Database.

### Working with the Time Dimension

Reporting change over time is a key element of the Management Reporting capability which has particular elements to consider.

The MRD is routinely updated on a weekly schedule. Each fact table includes a slice of data for each update and relationships with the DateDimension allow historic reporting.

When the Management Reporting Database is processed, any gap in the timeline for the fact tables will be populated by carrying forward the last known state.

To produce a single value for a given time period (e.g., reporting by month, year, etc.) you must choose whether to calculate an aggregate value for the period (e.g., Average, Max, Sum, etc.) or selecting an individual value. The DateDimension table includes flags denoting end of month, quarter and year, which can be used to select a single representative value for a time period. The correct way to determine the value for a time period will depend on the type of report and your business requirements.



# Oldest Compatible MRD Versions Matrix

As of Discovery Center version 4.7.0, there is no longer a requirement for the Management Reporting Database version to be the same as the Discovery Center version. Below is a matrix showing the oldest Management Reporting Database versions that are compatible with each Discovery Center version since this change was made in 4.7.0.

Table 21 Oldest Compatible MRD Versions

Discovery Center Version	Oldest Compatible Management Reporting Database Version
4.7.0	4.5.0
4.8.0	4.5.0
4.8.1	4.5.0
4.8.2	4.5.0
4.9.0	4.5.0
4.10.0	4.5.0
4.10.1	4.5.0
4.11.0	4.11.0
4.11.1	4.11.0
4.12.0	4.11.0
4.13.0	4.11.0
4.14.0	4.14.0
4.14.1	4.14.0
4.14.2	4.14.0
4.14.3	4.14.0
4.14.4	4.14.4
4.15.0	4.14.4
4.15.1	4.14.4
4.16.0	4.14.4
4.17.0	4.14.4



# Appendix 8: Connector Compatibility Summary

## Connector Capabilities Summary (I)

	Connector		
	File Share	SharePoint	Exchange
<b>Repository Information</b>			
Versions and variants	SMB + NFS with configuration	SP 2013, 2016, 2019, optimized for SPO	2010 onwards inc. Office365
API Notes	Windows API	SharePoint REST API	Exchange web services 2010 <b>Exchange On-Prem:</b> Remote PowerShell <b>Exchange Online:</b> MS Graph API
API Performance	Excellent	Good	Poor
Set Up Complexity	Low	Moderate	High
Connection and Authentication	NTLM Authentication on Windows	<b>SPO:</b> Using certificates for App based client credential authentication* <b>On-Prem:</b> Basic authentication	<b>Online:</b> Using tokens via App based ROPC authentication flow* <b>On-Prem:</b> Using NTLM <b>Not supported:</b> Kerberos authentication
<b>Discovery (skim)</b>			
Start locations	Server, share, folder	Site collection, site, library	Server, mailbox virtual group, mailbox
Structure	Server, share, folder structure	Site collection, site, document library, folder	Server, mailbox virtual group, mailbox, mailbox folder
Content objects	Files	Files in document libraries	Mail messages and attachments in personal and archive mailboxes
Read repository metadata	No	Named columns**	No
<b>Analysis</b>			
Content objects	Files	Files in document libraries	Mail messages and attachments
<b>Actions</b>			
Delete from	Yes	Yes	Yes
Quarantine from	Yes	Yes	Yes
Quarantine to	Yes	Yes	No
Migrate from	Yes	Yes	Yes
Migrate to	Yes	Yes	No
Label in Place	Yes	Yes	No
Write metadata to	Basic metadata only	Existing columns only	No



## Connector Capabilities Summary (II)

	Connector			
	OpenText Content Server	OpenText Content Server (Beta)	GSuite Google Drive	OneDrive
<b>Repository Information</b>				
Versions and variants	Version 10 onwards Customizations***	Version 10 onwards Customizations***		OneDrive connector currently prototype
API Notes	Web services via SeeUnity service	OTCS REST API	Google REST API	SharePoint web services 2007
API Performance	Poor	Moderate	Moderate	Moderate
Set Up Complexity	High	Moderate	Moderate	High
Connection and Authentication	Windows authentication via Content Server web services (must be enabled)	Windows authentication via Content Server web services (must be enabled)	Authorized service account via Google	Active Directory, native Office 365, ADFS
<b>Discovery (skim)</b>				
Start locations	Enterprise Workspace, Personal Workspaces, folder	Enterprise Workspace, folder	Drive, folder	One index per user OneDrive
Structure	Workspace, folder	Workspace, folder	Team drives, personal drives, folders	User OneDrive, folder
Content objects	File content types	File content types	Files	Files in folders
Read repository metadata	Named category fields	Named category fields	No	No
<b>Analysis</b>				
Content objects	File content types	File content types	Files	Files
<b>Actions</b>				
Delete from	Yes	Yes	Yes	Yes
Quarantine from	Yes	Yes	Yes	Yes
Quarantine to	Yes	Yes	Yes	Yes
Migrate from	Yes	Yes	Yes	Yes
Migrate to	Yes	Yes	Yes	Yes - cannot create folders
Label in Place	No	No	No	No
Write metadata to	Custom fields not supported	No	Basic metadata only	No



## Connector Capabilities Summary (III)

	Connector	
	Atlassian Confluence	Atlassian JIRA
<b>Repository Information</b>		
Versions and variants	6.12.1 onwards, on-premise and Atlassian cloud service	JIRA Core REST API 7.6.1 onwards, on-premise and Atlassian cloud service
API Notes	Atlassian web services	Atlassian web services
API Performance	Moderate	Moderate
Set Up Complexity	Moderate	Moderate
Connection and Authentication	<b>On-Prem:</b> Using basic authentication <b>Atlassian Cloud:</b> Using API tokens	<b>On-prem:</b> Using basic authentication <b>Atlassian Cloud:</b> Using API tokens
<b>Discovery (skim)</b>		
Start locations	Server, space, page blogs	Server, project type virtual group, alphabetical virtual group, project, issue
Structure	Server, spaces, pages, blogs	Server, project type virtual groups, alphabetical virtual groups, projects, issues, comments, worklogs, attachments
Content objects	Pages, blogs, comments and file attachments	Issues, worklogs, comments and file attachments
Read repository metadata	No	No
<b>Analysis</b>		
Content objects	Pages, blogs, comments and file attachments	Issues, worklogs, comments, attachments
<b>Actions</b>		
Delete from	No	No
Quarantine from	No	No
Quarantine to	No	No
Migrate from	Yes - cannot delete original	Yes - cannot delete original
Migrate to	No	No
Label in Place	No	No
Write metadata to	No	No

\* see Appendices of Discovery Center Installation Guide for configuration details for Microsoft Online repositories

\*\* see SharePoint Properties section for full details

\*\*\* pre-production deployment recommended to validate impact of repository customizations; case insensitive database presumed



# Appendix 9: Glossary

## Activity

Any work that is put on the Discovery Center Queue and then processed by the ActiveNav scheduler.

## AN Administrator

User responsible for defining and scheduling indexes for analysis. Organizations may have users that are both System Administrators and AN Administrators or who are both AN Administrators and Information Managers.

## Analysis

The process of extracting information from file content with advanced analysis techniques.

Requires access to file content (increased bandwidth requirement) and greater CPU usage than skim and file throughput is significantly slower.

## Analysis Queue

The Discovery Center can only carry out one Index analysis task at a time. If it is busy then any additional tasks that are due to be carried out at that time will be added to a queue and the first item in the queue will be started when Discovery Center becomes available.

## Analysis Schedule

The Analysis Schedule is the schedule of upcoming Index Analysis Tasks. When the date that a task is scheduled for is reached the task is added to the end of the Analysis Queue.

Tasks can either be added to the schedule when an Index is created or, after a task is completed, a check is made to see if the index should be kept up to date. If so, the task will be rescheduled after the appropriate interval.

## Area of Interest (AOI)

The AOI is a set of independent but related locations defined by an Information Manager to have some business-relevant meaning, for example, to define the location of information owned by a finance team. An AOI allows horizontal reporting as opposed to the strict vertical silos represented by Indexes and provides a bookmark for further exploration of information.

## Basic Metadata

Explicit Metadata found during skim that is stored in dedicated columns in the database schema for each document. This can optionally be published to the classification index for and important items have dedicated reports, for example Last Modified Date report or the File extension report.



## Calculated Fields

Named items of metadata that have been derived or assigned to a document because of the value of some Extracted Data or other Metadata Field value. Metadata Fields are available for migration to SharePoint and Reporting, and as Filters in reports, and can be used to support migration. Metadata Fields can be of the following types: Best Matching Rule, Any Matching Rules, First Value Classification or All Values Classification. Metadata Fields may be given default values, and may be single valued or multi-valued according to the selection configuration options.

## Central Server

A Discovery Center installation that has received content via import (manual or automated) from a remote server. Once a system has become a central server by virtue of performing an import we may not support further export of content.

## Chart

The visual presentation of a chosen dimension, for data extracted from an Index, AOI or Folder, using a specified Rule for selecting content

## Classification

A classification is a general structure used to group items together that have common attributes.

## Cluster

A collection of files which are duplicates of, or similar to, each other.

## Container

Any resource that can contain other resources, for example a folder on a file system or .zip archive file.

## Content Duplicate

Files that have identical content but different metadata due to certain automated updates by the containing repository or creating application. For example, SharePoint will update document creation and modified dates inside the content of a Microsoft Office document when it is saved in a SharePoint library. These changes make the files different (non-duplicate), but Discovery Center can ignore these changes to identify Content Duplicate documents.

## Content Regular Expression Matches

A piece of document text that matches a regular expression, or capture group within a regular expression that has been applied to the textual content of the document.

## Coverage

Effectiveness of classification structure in matching existing files to new nodes.



## Crawl

Generic term used to describe the creation and population of indexes or network maps by the Discovery Center. Encompasses network discovery, skim and analysis.

## Discovery Center

All of the web application parts of Discovery Center including the administration and reporting pages.

## Diversity (a type of field score)

The number of unique values for a file for a specific field. Only up to 1,000 unique values will be recorded for any match and so 1,000 is the maximum possible diversity for a field.

## Duplicate Files

Files identical at bit level. In practical terms, this means files that are exact copies of each other. (Also see **Content Duplicate**).

## Excluded Location

Users in the System Administrator role can configure locations to be excluded from Discovery Center analysis using the Network Map. These locations and any child content or locations will not be shown in the Network Maps anywhere in the UI except the Network Map tab visible to the System Administrators.

## Explicit Metadata

Existing descriptive properties of a file.

## Extracted Data

Named items of Metadata that have been directly extracted from a document (like a title string), or can be directly measured (like a file size). Approaches to extract data include skimming, thematic analysis, entity extraction, pattern matching.

## Facet

Searchable data extracted during indexing based on Basic Metadata, Folder names, Themes or Calculated Fields.

## Feature Pack

A license consists of one or more Feature Packs and these provide access to certain functions in Discovery Center. Each Pack collects together logically associated features such as text analysis or file deletion. A license should include the specific feature packs required for a *Solution*. Unlicensed features within Discovery Center are hidden, greyed out or the capability is blocked.





### Field score

A number representing the value of a metric for a field for a particular file or container.

### Field Source

The destination location used to provide an example of available metadata fields for mapping.

### File Extension

Any sequence of text after and including the final '.' in a file name. (Strictly the '.' is considered to be part of the extension, and if the extension text is more than 40 characters then it is not considered to be an extension).

### File Format

The description of a structure for a file that can be read by one or more applications, perhaps specific to the application that created the file. Oracle's OutsideIn filters provide details for file format for analyzed files.

### File Path Regular Expression Matches

A part of a file full path name that matches a regular expression, or capture group within a regular expression that has been applied to the file full path name.

### File System Properties

Information read from the file system about each document including document Created Date, Last Modified Date and File Size. There are dedicated reports available for many of these attributes in the Reporting UI. This information is part of the Basic Metadata available for use in Classification.

### First Value Field

The first value found for one or more named pieces of Extracted Data from a defined sequence of Extracted Data items, with an optional default value.

This type of Field could be used to make a one-to-one mapping from a selection of Extracted Data for a document to a named piece of Field Metadata.

### Filter

Criteria for selecting content from within a chosen Index, AOI or Folder. For example, File size > 1Gb, File Age > 2 years, root location or AOI, or Field Value.

### Folder

A type of container created within a computer storage environment (such as a file share or Microsoft SharePoint) for the organization and grouping of related files.



## Group

Refers to a Windows Group of Users.

## Heat

Presentation of Matched File Count or Intensity as a color range overlaid on a Container view.

## Index

An analysis run from the Discovery Center to identify duplication, themes and to extract metadata which can be used by the Discovery Center to deliver the data migration and cleansing methodology. An Index is the basic element of analysis, defined by the index configuration that has been applied to it.

## Index Configuration

The settings used during Index processing. The settings include depth of analysis, calculated fields and other options. Any index configuration can be used to control one or more indexes.

## Index Ignored Location

A container that will be ignored during discovery. No files or sub-containers will be added to the index, but the ignored location itself will be visible on the Network Map.

## Implicit Metadata

Properties deduced from the content of a file or the way that it is stored or used.

## Index Security

An index setting that determines which Discovery Center users will be able to see file level detail (file lists and file properties) of the index via the reporting UI.

## Information Manager

User responsible for interrogating the system to generate reports, identify policy violations and remove ROT.

## Intensity

*(a type of field score)* The total number of matches (possibly repeated values) for a file for a specific field. Only up to 1,000 unique values are considered in any file, but this value might be more than 1,000 if values are repeated.

For Containers, Intensity is the sum of the Intensity values for all of the contained files.



## Keyword

A type of text extraction rule that defines a text pattern to record preferred terms of synonym matches in a file, and count the number of matches found. Only the preferred terms will be recorded when synonyms are matched.

## Location

An item in the Network Map, as added in the Network Map tab, or select as a start location for an Index or View.

## Management Reporting Database

If the appropriate license is applied, the Management Reporting Database accumulates historical snapshots of the primary database state to allow for monitoring of information management metrics such as key calculated fields, storage costs, activities performed.

## Matched File Count

*(a type of field score)* The number of files that have one or more values for a selected field.

## Metadata

Any extracted (explicit) or calculated (implicit) attribute of a document.

## Network Credentials

An index setting that determines which user credentials will be used to access file content during discovery, analysis and actions.

## Network Map

A tree presentation of content repositories and their child-content down to container (folder or equivalent) level, and displayed in various parts of the Discovery Center application. Users in the System Administrator role can exclude selected containers from presentation in the Discovery Center and from any analysis by using the Network Map tab.

The Network Map tab also includes high level properties of repositories and folders.

## Node

An element of a classification hierarchy to which files are assigned based on the specified node rules.

## Node Rule

Query in Classification Workbench (formerly Classification Designer) constructed using search Facets and discovered terms or properties.

## Orphaned file

File that Classification Workbench (formerly Classification Designer) cannot assign to a node under existing node rules.



#### \* Other

Indicates that 2 or more values have been aggregated into a single section of a chart. This happens when showing charts with more than 30 columns, and when showing Container reports with segments that are less than 1/200th of the total data presented.

#### Remote Server

A Discovery Center installation that has provided content via export (manual or automated).

#### Report

A group of one or more charts and associated text.

#### Reporting Database

Primary reporting database that allows the generation of interactive reports of the currently indexed content.

#### Resource Rank

The relative order of files or containers taking into account the file metadata and any field that is currently under consideration by the user so that the most relevant files are selected for display.

#### Reviewer

The user receiving the final output from Discovery Center in the form of reports and Work Packages.

#### Roles

Discovery Center is designed around the following key roles:

- System Administrator
- AN Administrator
- Information Manager
- Reviewer

Access to functions is controlled by Role.

#### ROT

Redundant, Obsolete or Trivial: describes the types of documents to be cleansed from the file share.

#### Sample Files

A representative (and not necessarily complete) selection of files that have been included in the currently viewed report. The selection of which files to include has been made using our Resource Rank algorithm.



## Saved View

The Report Type, Filter settings, Chart Axis settings and Area of Interest or Network Location selected by a user. This is a dynamic view of a particular chart or file view (or report). If a Saved View is loaded after it was initially created, the chart may look different if the underlying data has changed.

## Similar Files

Files that are not identical but share similar textual content as determined by Discovery Center thematic fingerprints.

## Skim

The process of collecting together structure and property information without analysis of file content.

## Solution

A licensed solution includes all the necessary Feature Packs to fulfill a particular objective: Intelligent Migration, Content Compliance and/or Content Governance.

## Storage Tier

Specifies a class of storage that can be assigned to locations on the Network Map. Specifying a cost for 1 GB of content held in the specified storage class allows the system to estimate the storage cost associated with specified Storage Tiers.

## Summary

A short extract of text composed from sentences found inside a document that best represent the content of the document. The number of sentences included in a summary is configurable.

## Thematic Metadata

Metadata that has been extracted from a (textual) natural language document using Discovery Center thematic analysis: Titles, Summaries, Themes and Forced Themes.

## Title

A string intended by a document author to represent the title of a document. If the author has not assigned any title attribute (or such an attribute is considered unreliable) then the Thematic Analysis has to deduce the correct title string from document formatting.

## Themes

Words and phrases found in a document that best represent the key themes of the document content.



### Unclassified theme

A theme not used to classify any file within the current structure design.

### User

Refers to an individual Windows user.



Copyright © 2022 Data Discovery, Limited. All Rights Reserved

ActiveNav® is a registered trademark of Data Discovery Solutions Ltd in the United States and other countries. All trademarks used herein are the property of their respective owners.

ActiveNav believes that information in this publication is accurate as of its publication date. The information is subject to change without notice.

The information in this publication is provided “as is.” ActiveNav make no representation or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantable or fitness for a particular purpose.

