Table of Contents

1. Overview ........................................................................................................................................ 9
   1.1. Features and Benefits .................................................................................................................. 9
   1.2. How It Works .............................................................................................................................. 10
2. Deployment ...................................................................................................................................... 12
   2.1. Supported Data Sources .............................................................................................................. 21
   2.2. Deployment Planning .................................................................................................................. 21
       2.2.1. NDC Server ....................................................................................................................... 21
       2.2.2. Data Storages and Sizing .................................................................................................... 22
           2.2.2.1. Scalability and Performance ....................................................................................... 23
           2.2.2.2. Recommendations on SQL Database Maintenance .................................................... 23
       2.2.3. Scalability and Performance .............................................................................................. 24
           2.2.3.1. Example: Mid-Size Data Environment ........................................................................ 25
           2.2.3.2. Example: Large-Size Environment .............................................................................. 28
   2.3. Requirements to Install Netwrix Data Classification ................................................................. 32
       2.3.1. Hardware Requirements ..................................................................................................... 33
           2.3.1.1. Netwrix Data Classification Server .......................................................................... 33
           2.3.1.2. SQL Server ................................................................................................................. 33
           2.3.1.3. Network Access ......................................................................................................... 34
           2.3.1.4. Configuring NDC Servers Cluster and Load Balancing with DQS Mode ................. 34
       2.3.2. Software Requirements ...................................................................................................... 38
       2.3.3. Accounts and Required Permissions .................................................................................... 40
   2.4. Configure NDC Database ........................................................................................................... 42
   2.5. Install Netwrix Data Classification ............................................................................................ 43
   2.6. Upgrade to the Latest Version ................................................................................................... 44
       2.6.1. Take Preparatory Steps ..................................................................................................... 44
       2.6.2. Considerations and Limitations ......................................................................................... 44
   2.7. Configuring NDC Servers Cluster and Load Balancing with DQS Mode .................................. 45
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.7.1.</td>
<td>Applying DQS Mode</td>
<td>45</td>
</tr>
<tr>
<td>2.8.</td>
<td>Configure IT Infrastructure</td>
<td>49</td>
</tr>
<tr>
<td>2.8.1.</td>
<td>Configure Microsoft Exchange for Crawling and Classification</td>
<td>50</td>
</tr>
<tr>
<td>2.8.2.</td>
<td>Configure NFS File Share for Crawling</td>
<td>53</td>
</tr>
<tr>
<td>2.8.3.</td>
<td>Configure G Suite for Crawling</td>
<td>53</td>
</tr>
<tr>
<td>2.9.</td>
<td>Initial Product Configuration</td>
<td>56</td>
</tr>
<tr>
<td>2.9.1.</td>
<td>Select Processing Mode</td>
<td>56</td>
</tr>
<tr>
<td>2.9.2.</td>
<td>Processing Settings</td>
<td>57</td>
</tr>
<tr>
<td>2.9.3.</td>
<td>Add Taxonomy</td>
<td>57</td>
</tr>
<tr>
<td>2.9.4.</td>
<td>Review Your Configuration</td>
<td>58</td>
</tr>
<tr>
<td>3.</td>
<td>Security (Users)</td>
<td>59</td>
</tr>
<tr>
<td>3.1.</td>
<td>Secure Netwrix Data Classification</td>
<td>59</td>
</tr>
<tr>
<td>3.2.</td>
<td>User Management</td>
<td>61</td>
</tr>
<tr>
<td>3.3.</td>
<td>Password Manager</td>
<td>64</td>
</tr>
<tr>
<td>3.4.</td>
<td>Web Service Security</td>
<td>65</td>
</tr>
<tr>
<td>4.</td>
<td>Content Sources</td>
<td>66</td>
</tr>
<tr>
<td>4.1.</td>
<td>Add a Content Source</td>
<td>66</td>
</tr>
<tr>
<td>4.1.1.</td>
<td>Database</td>
<td>67</td>
</tr>
<tr>
<td>4.1.2.</td>
<td>Exchange Mailbox</td>
<td>68</td>
</tr>
<tr>
<td>4.1.3.</td>
<td>Exchange Server</td>
<td>69</td>
</tr>
<tr>
<td>4.1.4.</td>
<td>File System</td>
<td>70</td>
</tr>
<tr>
<td>4.1.4.1.</td>
<td>Add Folder source</td>
<td>70</td>
</tr>
<tr>
<td>4.1.4.2.</td>
<td>Add Files source</td>
<td>72</td>
</tr>
<tr>
<td>4.1.5.</td>
<td>Add Google Drive Source</td>
<td>73</td>
</tr>
<tr>
<td>4.1.6.</td>
<td>Outlook Mail Archive</td>
<td>75</td>
</tr>
<tr>
<td>4.1.7.</td>
<td>SharePoint</td>
<td>75</td>
</tr>
<tr>
<td>4.1.8.</td>
<td>SharePoint Online</td>
<td>76</td>
</tr>
<tr>
<td>4.2.</td>
<td>Narrow Data Collection Scope</td>
<td>77</td>
</tr>
<tr>
<td>4.3.</td>
<td>Use Tagging (optional)</td>
<td>77</td>
</tr>
<tr>
<td>4.4.</td>
<td>Manage Sources and Control Data Processing</td>
<td>78</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>4.4.1. Modify Source Settings</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>4.4.2. Set up granular processing and tagging for Database</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>4.4.3. Set up exclusions and tagging for Exchange</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>4.4.4. Set up filters and tagging for File System</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>4.4.4.1. Configure Inclusions</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>4.4.4.2. Configure Exclusions</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>4.4.4.3. Configure Tagging</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>4.4.5. Set up exclusions and tagging for Google Drive</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>4.4.6. Set up processing options for SharePoint</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>4.4.7. Set up processing options for SharePoint Online Tenancy</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>4.5. View Results</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>4.5.1. Data Processing Statistics</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>4.5.2. Content Crawling and Classification Results</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>5. Taxonomies</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>5.1. What are Taxonomies?</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>5.2. Built-in Taxonomies Overview</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>5.2.1. Core Taxonomies</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>5.2.2. Derived Taxonomies</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>5.3. Taxonomy Settings</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td>5.3.1. Taxonomy Settings Levels</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td>5.3.2. Labels</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>5.3.2.1. SharePoint Labels</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>5.3.2.2. O365 Labels</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>5.3.2.3. Help</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>5.4. Add a Taxonomy</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>5.5. Manage Taxonomies</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>5.5.1. Managing Term Sets</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>5.5.2. Multi-User Environments</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>5.6. Search and Filter Taxonomies</td>
<td>115</td>
<td></td>
</tr>
<tr>
<td>5.7. Classification Rules (Clues)</td>
<td>118</td>
<td></td>
</tr>
</tbody>
</table>
5.7.1. Predefined Classification Rules ................................................................. 118
5.7.2. Working with Clues .................................................................................. 120
5.7.3. Documents count .................................................................................... 120
5.7.4. Suggested Clues ...................................................................................... 121
5.7.5. Types of Clues ......................................................................................... 122
5.7.6. Adding a Clue .......................................................................................... 131
  5.7.6.1. Clue Body ............................................................................................ 131
  5.7.6.2. Score .................................................................................................. 132
  5.7.6.3. Mandatory Clues ................................................................................ 133
  5.7.6.4. Using the Local Option ....................................................................... 133
  5.7.6.5. Using Synonyms (SQL taxonomies only) ............................................. 134
5.7.7. Manage Clues ........................................................................................... 134
  5.7.7.1. Bulk Edit ............................................................................................. 134
  5.7.7.2. Bulk Import ......................................................................................... 136
5.7.8. Search Documents by Clue ..................................................................... 136
5.7.9. Browse .................................................................................................... 138
5.7.10. Export Search Results ........................................................................... 139
5.8. Suggestions .................................................................................................. 140
5.9. Working Set ................................................................................................ 141
5.10. Related ...................................................................................................... 142
5.11. Additional Configuration .......................................................................... 142
6. Workflows ....................................................................................................... 145
  6.1. Understanding Workflows ........................................................................ 145
  6.2. Managing Workflows ................................................................................ 145
    6.2.1. Create a Workflow using Add Workflow Wizard .................................. 148
      6.2.1.1. Step 1. Select Content Type ............................................................ 149
      6.2.1.2. Step 2. Select Action ..................................................................... 150
      6.2.1.3. Step 3. Specify Conditions for Processing ..................................... 151
      6.2.1.4. Step 4. Enter Name and Review Settings ....................................... 155
    6.2.2. Configure a Workflow using Advanced dialog .................................... 156
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2.7. Additional Configuration Settings</td>
<td>195</td>
</tr>
<tr>
<td>7.2.8. Configuration Backup</td>
<td>198</td>
</tr>
<tr>
<td>7.3. Review Dashboards</td>
<td>200</td>
</tr>
<tr>
<td>7.3.1. System Health</td>
<td>201</td>
</tr>
<tr>
<td>7.3.2. Netwrix Data Classification Service Viewer</td>
<td>201</td>
</tr>
<tr>
<td>8. Reporting Capabilities</td>
<td>202</td>
</tr>
<tr>
<td>8.1. Content Distribution</td>
<td>203</td>
</tr>
<tr>
<td>8.2. Review Built-in Reports</td>
<td>203</td>
</tr>
<tr>
<td>8.3. Types of Reports</td>
<td>206</td>
</tr>
</tbody>
</table>
Legal Notice

The information in this publication is furnished for information use only, and does not constitute a commitment from Netwrix Corporation of any features or functions, as this publication may describe features or functionality not applicable to the product release or version you are using. Netwrix makes no representations or warranties about the Software beyond what is provided in the License Agreement. Netwrix Corporation assumes no responsibility or liability for the accuracy of the information presented, which is subject to change without notice. If you believe there is an error in this publication, please report it to us in writing.

Netwrix is a registered trademark of Netwrix Corporation. The Netwrix logo and all other Netwrix product or service names and slogans are registered trademarks or trademarks of Netwrix Corporation. Microsoft, Active Directory, Exchange, Exchange Online, Office 365, SharePoint, SQL Server, Windows, and Windows Server are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. All other trademarks and registered trademarks are property of their respective owners.

Disclaimers

This document may contain information regarding the use and installation of non-Netwrix products. Please note that this information is provided as a courtesy to assist you. While Netwrix tries to ensure that this information accurately reflects the information provided by the supplier, please refer to the materials provided with any non-Netwrix product and contact the supplier for confirmation. Netwrix Corporation assumes no responsibility or liability for incorrect or incomplete information provided about non-Netwrix products.

© 2019 Netwrix Corporation.

All rights reserved.
1. Overview

1.1. Features and Benefits

Netwrix Data Classification is a platform that identifies data that’s important for your organization and enables you to reduce risk and unleash the true value of this data.

Powered by unique compound term processing technology, it enriches your enterprise content with accurate and consistent metadata empowering you to work with data more confidently. By seeing which data is valuable, you can organize it in a way that promotes productivity and collaboration. By knowing where sensitive or regulated data is, you can reduce the risk of breaches and satisfy security and privacy requirements with less effort and expense. And by locating and getting rid of redundant and obsolete data, you can reduce storage and management costs.

Netwrix Data Classification includes applications for Windows File Servers, Nutanix Files, Dell EMC, NetApp, and SharePoint, Office 365, Exchange, SQL Server, Oracle Database, Box, Google Drive, MySQL and PostgreSQL. The platform provides a single panoramic view of your enterprise content, whether it’s located in structured or unstructured data stores, on premises or in the cloud.

Major benefits:

- Identify sensitive information and reduce its exposure
- Improve employee productivity and decision making
- Reduce costs and risks by getting rid of unneeded data
- Meet privacy and compliance requirements for information governance
- Respond to legal requests without putting your business on hold

Netwrix Discover is a data discovery and classification tool based on the Netwrix Data Classification platform. The tool is designed to help managed service providers bring value to existing customers and attract new clients by identifying their sensitive data and its location. This initial assessment enables you to start the conversation about what security controls your customers and prospects have in place and at what level they would like to protect their data so you can offer managed security services.

Unstructured data security

Automatically identify sensitive unstructured data on customers’ file servers and SharePoint (including SharePoint Online) as well as their repositories with the highest concentration of these critical files.

Predefined taxonomies

Kick off your discovery with out-of-the-box classification rules that identify data regulated by PCI DSS, HIPAA, GDPR, CCPA and other regulations.

Transparent classification rules and results
Demonstrate to the customer why files were classified as they were. Deliver accurate results instead of wasting customers’ time sifting through false positives.

**Non-intrusive deployment**

Leverage a tool that operates in agentless mode and does not interfere with your customers’ file and SharePoint systems.

### 1.2. How It Works

1. The user enters data sources using the **administrative web console**.
2. The configured data sources are added in the **NDC SQL database**.
3. The **NDC Collector** service crawls data files in each data source, converts documents into plain text and populates file metadata in the **NDC SQL database**.
4. The **NDC Indexer** service builds and maintains a full-text search index (**NDC Index**) based on the content and metadata of the collected files.
5. The **NDC Classifier** service performs data classification by matching collected files against installed taxonomies (e.g., Netwrix compliance taxonomies).
6. If **Data Tagging** is enabled, the assigned classification labels are written to the custom metadata columns for supported document types.
7. If **Remediation Workflows** are enabled, the configured workflows are run on documents that meet the workflow conditions.

**Legal Notice**

The information in this publication is furnished for information use only, and does not constitute a commitment from Netwrix Corporation of any features or functions, as this publication may describe features or functionality not applicable to the product release or version you are using. Netwrix makes no representations or warranties about the Software beyond what is provided in the License Agreement. Netwrix Corporation assumes no responsibility or liability for the accuracy of the information presented, which is subject to change without notice. If you believe there is an error in this publication, please report it to us in writing.

Netwrix is a registered trademark of Netwrix Corporation. The Netwrix logo and all other Netwrix product or service names and slogans are registered trademarks or trademarks of Netwrix Corporation. Microsoft, Active Directory, Exchange, Exchange Online, Office 365, SharePoint, SQL Server, Windows, and Windows Server are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. All other trademarks and registered trademarks are property of their respective owners.

**Disclaimers**

This document may contain information regarding the use and installation of non-Netwrix products. Please note that this information is provided as a courtesy to assist you. While Netwrix tries to ensure that this information accurately reflects the information provided by the supplier, please refer to the materials provided with any non-Netwrix product and contact the supplier for confirmation. Netwrix Corporation
assumes no responsibility or liability for incorrect or incomplete information provided about non-Netwrix products.

© 2019 Netwrix Corporation.

All rights reserved.
2. Deployment

Netwrix Data Classification

User Guide

Version: 5.5.2
# Table of Contents

1. Overview .......................................................................................................................... 9
   1.1. Features and Benefits ................................................................................................. 9
   1.2. How It Works ............................................................................................................. 10
2. Deployment .......................................................................................................................... 12
   2.1. Supported Data Sources ......................................................................................... 21
   2.2. Deployment Planning ............................................................................................... 21
       2.2.1. NDC Server ......................................................................................................... 21
       2.2.2. Data Storages and Sizing .................................................................................. 22
           2.2.2.1. Scalability and Performance ...................................................................... 23
           2.2.2.2. Recommendations on SQL Database Maintenance .................................. 23
       2.2.3. Scalability and Performance .............................................................................. 24
           2.2.3.1. Example: Mid-Size Data Environment ...................................................... 25
           2.2.3.2. Example: Large-Size Environment .......................................................... 28
   2.3. Requirements to Install Netwrix Data Classification .................................................. 32
       2.3.1. Hardware Requirements ................................................................................... 33
           2.3.1.1. Netwrix Data Classification Server ............................................................. 33
           2.3.1.2. SQL Server .................................................................................................. 33
           2.3.1.3. Network Access ........................................................................................ 34
           2.3.1.4. Configuring NDC Servers Cluster and Load Balancing with DQS Mode .... 34
       2.3.2. Software Requirements ...................................................................................... 38
       2.3.3. Accounts and Required Permissions .................................................................. 40
   2.4. Configure NDC Database ............................................................................................ 42
   2.5. Install Netwrix Data Classification ............................................................................. 43
   2.6. Upgrade to the Latest Version ..................................................................................... 44
       2.6.1. Take Preparatory Steps ...................................................................................... 44
       2.6.2. Considerations and Limitations .......................................................................... 44
   2.7. Configuring NDC Servers Cluster and Load Balancing with DQS Mode .................. 45
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.7.1. Predefined Classification Rules</td>
<td>118</td>
</tr>
<tr>
<td>5.7.2. Working with Clues</td>
<td>120</td>
</tr>
<tr>
<td>5.7.3. Documents count</td>
<td>120</td>
</tr>
<tr>
<td>5.7.4. Suggested Clues</td>
<td>121</td>
</tr>
<tr>
<td>5.7.5. Types of Clues</td>
<td>122</td>
</tr>
<tr>
<td>5.7.6. Adding a Clue</td>
<td>131</td>
</tr>
<tr>
<td>5.7.6.1. Clue Body</td>
<td>131</td>
</tr>
<tr>
<td>5.7.6.2. Score</td>
<td>132</td>
</tr>
<tr>
<td>5.7.6.3. Mandatory Clues</td>
<td>133</td>
</tr>
<tr>
<td>5.7.6.4. Using the Local Option</td>
<td>133</td>
</tr>
<tr>
<td>5.7.6.5. Using Synonyms (SQL taxonomies only)</td>
<td>134</td>
</tr>
<tr>
<td>5.7.7. Manage Clues</td>
<td>134</td>
</tr>
<tr>
<td>5.7.7.1. Bulk Edit</td>
<td>134</td>
</tr>
<tr>
<td>5.7.7.2. Bulk Import</td>
<td>136</td>
</tr>
<tr>
<td>5.7.8. Search Documents by Clue</td>
<td>136</td>
</tr>
<tr>
<td>5.7.9. Browse</td>
<td>138</td>
</tr>
<tr>
<td>5.7.10. Export Search Results</td>
<td>139</td>
</tr>
<tr>
<td>5.8. Suggestions</td>
<td>140</td>
</tr>
<tr>
<td>5.9. Working Set</td>
<td>141</td>
</tr>
<tr>
<td>5.10. Related</td>
<td>142</td>
</tr>
<tr>
<td>5.11. Additional Configuration</td>
<td>142</td>
</tr>
<tr>
<td>6. Workflows</td>
<td>145</td>
</tr>
<tr>
<td>6.1. Understanding Workflows</td>
<td>145</td>
</tr>
<tr>
<td>6.2. Managing Workflows</td>
<td>145</td>
</tr>
<tr>
<td>6.2.1. Create a Workflow using Add Workflow Wizard</td>
<td>148</td>
</tr>
<tr>
<td>6.2.1.1. Step 1. Select Content Type</td>
<td>149</td>
</tr>
<tr>
<td>6.2.1.2. Step 2. Select Action</td>
<td>150</td>
</tr>
<tr>
<td>6.2.1.3. Step 3. Specify Conditions for Processing</td>
<td>151</td>
</tr>
<tr>
<td>6.2.1.4. Step 4. Enter Name and Review Settings</td>
<td>155</td>
</tr>
<tr>
<td>6.2.2. Configure a Workflow using Advanced dialog</td>
<td>156</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>7.2.7. Additional Configuration Settings</td>
<td>195</td>
</tr>
<tr>
<td>7.2.8. Configuration Backup</td>
<td>198</td>
</tr>
<tr>
<td>7.3. Review Dashboards</td>
<td>200</td>
</tr>
<tr>
<td>7.3.1. System Health</td>
<td>201</td>
</tr>
<tr>
<td>7.3.2. Netwrix Data Classification Service Viewer</td>
<td>201</td>
</tr>
<tr>
<td>8. Reporting Capabilities</td>
<td>202</td>
</tr>
<tr>
<td>8.1. Content Distribution</td>
<td>203</td>
</tr>
<tr>
<td>8.2. Review Built-in Reports</td>
<td>203</td>
</tr>
<tr>
<td>8.3. Types of Reports</td>
<td>206</td>
</tr>
</tbody>
</table>
Legal Notice

The information in this publication is furnished for information use only, and does not constitute a commitment from Netwrix Corporation of any features or functions, as this publication may describe features or functionality not applicable to the product release or version you are using. Netwrix makes no representations or warranties about the Software beyond what is provided in the License Agreement. Netwrix Corporation assumes no responsibility or liability for the accuracy of the information presented, which is subject to change without notice. If you believe there is an error in this publication, please report it to us in writing.

Netwrix is a registered trademark of Netwrix Corporation. The Netwrix logo and all other Netwrix product or service names and slogans are registered trademarks or trademarks of Netwrix Corporation. Microsoft, Active Directory, Exchange, Exchange Online, Office 365, SharePoint, SQL Server, Windows, and Windows Server are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. All other trademarks and registered trademarks are property of their respective owners.

Disclaimers

This document may contain information regarding the use and installation of non-Netwrix products. Please note that this information is provided as a courtesy to assist you. While Netwrix tries to ensure that this information accurately reflects the information provided by the supplier, please refer to the materials provided with any non-Netwrix product and contact the supplier for confirmation. Netwrix Corporation assumes no responsibility or liability for incorrect or incomplete information provided about non-Netwrix products.

© 2019 Netwrix Corporation.

All rights reserved.
2.1. Supported Data Sources

The table below lists systems that can be crawled with Netwrix Data Classification:

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Supported Versions</th>
</tr>
</thead>
<tbody>
<tr>
<td>File System</td>
<td>• CIFS/SMB (Preferred)</td>
</tr>
<tr>
<td></td>
<td>• NFS</td>
</tr>
<tr>
<td>SharePoint, SharePoint Online, OneDrive for Business</td>
<td>• 2010 and above</td>
</tr>
<tr>
<td>Database</td>
<td>• Microsoft SQL Server 2008 and above</td>
</tr>
<tr>
<td></td>
<td>• Oracle 10g and above</td>
</tr>
<tr>
<td>Exchange</td>
<td>• Exchange Server 2010 and above</td>
</tr>
<tr>
<td></td>
<td>• Exchange Online</td>
</tr>
<tr>
<td>Google Drive</td>
<td>• N/A</td>
</tr>
<tr>
<td>Outlook Mail Archive</td>
<td>• Outlook 2010 and above</td>
</tr>
</tbody>
</table>

2.2. Deployment Planning

This section provides recommendations and considerations for Netwrix Data Classification deployment planning. Review these recommendations and choose the most suitable deployment scenario and possible options depending on the IT infrastructure and data sources you are going to process.

In this section:

- NDC Server
- Data Storages and Sizing
- Scalability and Performance

2.2.1. NDC Server

Netwrix Data Classification Server can be deployed on a physical server or on a virtual machine in the virtualized environment on VMware or Microsoft Hyper-V platform.

When planning for NDC Server, consider a significant CPU load during data processing. Thus, installing NDC Server on a highly-loaded production machine is not recommended. For more information, refer to Hardware Requirements.
Web-based client (management console) is always installed together with the NDC Server, so the IIS server role must be enabled on the target machine. For more information, refer to Software Requirements.

NOTE: For evaluation and PoC purposes, Netwrix provides a virtual appliance — a virtual machine image with pre-installed Netwrix Data Classification on Generalized Windows Server 2016 (180-day evaluation version) and Microsoft SQL Server 2017 Express. For details, see Requirements to Deploy Virtual Appliance.

Remember that for production environments, your NDC Server and database server must meet the Requirements to Install Netwrix Data Classification. Virtual appliance configuration is insufficient for production and is not recommended for that purpose.

To balance the load while indexing and classifying data in the large-size and extra-large environments (i.e. with over 8-10 mln objects to process), it is strongly recommended to deploy several NDC Servers and configure Distributed Query Server mode for them. See Configuring NDC Servers Cluster and Load Balancing with DQS Mode.

2.2.2. Data Storages and Sizing

Netwrix Data Classification utilizes two data storages:

- NDC SQL database — SQL Server database that stores product configuration and metadata for the data sources.
- NDC Index — a full-text search index that comprises a set of files in the proprietary format (.CSE).

2.2.2.0.1. NDC SQL database

Make sure you have NDC Server and NDC SQL database deployed on different machines.

It is recommended to create the NDC SQL database on a dedicated SQL Server instance.

- Minimal requirement is SQL Server 2008 R2 Standard Edition.
- Estimate required disk space assuming 10 - 12 KB per indexed object. For example, for 5,000,000 objects, the database size will be approximately 50 GB. Therefore, SQL Server Express edition will be only suitable for evaluation and PoC environments (up to 1,000,000 documents to process).
- If configuring database settings via SQL Server Management Studio, you will need to set Autogrowth / Maxsize values for the PRIMARY database files as follows:
  - File growth: 128 MB - recommended value for small to medium environment, 512 MB - for large environment, i.e. if planning to index data sources containing 1,000,000+ objects.
  - Maximum File Size - select Unlimited.
- Make sure that the Recovery model for this database is set to Simple. Do not change the recovery model — to avoid log files growth.

See also Recommendations on SQL Database Maintenance.
2.2.0.2. NDC Index

Required disk space for the **NDC Index** file storage will depend, in particular, on the data processing mode you plan to use (*No Index*, *Keyword* or *Compound Term*).

As a rule of thumb, required space can be calculated as 35% of data you plan to be indexed. For example, if you have 45 GB of files, they will require up to 15 GB for the **NDC Index** files.

### 2.2.2.1. Scalability and Performance

Scalability and performance testing revealed that based on the number of objects to classify, the environments can be ranged as follows:

<table>
<thead>
<tr>
<th>Number of objects to classify</th>
<th>Environment</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 500,000</td>
<td>Proof-of-concept and small-size environment</td>
<td></td>
</tr>
<tr>
<td>Up to 8,000,000</td>
<td>Mid-size environment</td>
<td></td>
</tr>
<tr>
<td>Up to 32,000,000</td>
<td>Large-size environment</td>
<td></td>
</tr>
<tr>
<td>More than 32,000,000</td>
<td>Extra-large environment</td>
<td>System architect’s assistance is required for deployment planning requires</td>
</tr>
</tbody>
</table>

The following sections describe related deployment scenarios and provide examples for resource planning:

**Example: Mid-Size Data Environment**

**Example: Large-Size Environment**

You can use these examples to estimate hardware requirements and plan for scalability of your Data Classification deployment.

Again, consider that for the large-size and extra-large environments, it is strongly recommended to configure a cluster of several NDC Servers and apply DQS mode to these clustered servers. See [Configuring NDC Servers Cluster and Load Balancing with DQS Mode](#) for details.

### 2.2.2.2. Recommendations on SQL Database Maintenance

Netwrix Data Classification uses SQL Server database as a storage for file metadata prepopulated by **NDC Collector** service. If you are going to crawl more than 1M of objects, you need to pay attention to SQL Server database maintenance procedures, especially during initial collection period. You or your database administrator can perform these tasks according to your company’s internal policies, if any, or follow the recommendations below.
To ensure data integrity and performance, maintenance operations recommended by Microsoft should be performed for your NDC SQL database once a day, putting more focus on the Pages table.

To maintain your SQL database

IMPORTANT! Stop all Netwrix Data Classification services before you start the maintenance procedures. If you are using DQS (Distributed Query Server) mode, you need to stop all services on all instances of Netwrix Data Classification. You can stop and start the services, using the Netwrix Data Classification: Service Viewer tool.

Do the following:

1. On the computer where Netwrix Data Classification is installed, start the Netwrix Data Classification Service Viewer tool. Select Stop next to each service.
2. Start Microsoft SQL Management Studio and connect to the SQL Server instance hosting NDC SQL database.
3. Right-click the NDC SQL database and select Reports → Standard Reports → Index Physical Statistics report.
4. Based on the report data, perform the recommended operations (Rebuild or Reorganize) for the indexes of the certain tables. For details, see this Microsoft article: Reorganize and rebuild indexes

NOTE: The following indexes do not influence database performance during the initial data crawling, so you can skip them when performing the initial maintenance procedure:

  - Checksum
  - IdxPagesFileChecksum
  - IdxPagesTextChecksum
  - DocumentChecksum

After the initial crawling is completed, you can include these indexes in the standard daily database maintenance procedure.

2.2.3. Scalability and Performance

Scalability and performance testing revealed that based on the number of objects to classify, the environments can be ranged as follows:

<table>
<thead>
<tr>
<th>Number of objects to classify</th>
<th>Environment</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 500,000</td>
<td>Proof-of-concept and</td>
<td>small-size environment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up to 8,000,000</td>
<td>Mid-size environment</td>
<td></td>
</tr>
</tbody>
</table>
2. Deployment

<table>
<thead>
<tr>
<th>Number of objects to classify</th>
<th>Environment</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 32,000,000</td>
<td>Large-size environment</td>
<td></td>
</tr>
<tr>
<td>More than 32,000,000</td>
<td>Extra-large environment</td>
<td>System architect’s assistance is required for deployment planning in such environments.</td>
</tr>
</tbody>
</table>

The following sections describe related deployment scenarios and provide examples for resource planning:

- Example: Mid-Size Data Environment
- Example: Large-Size Environment

You can use these examples to estimate hardware requirements and plan for scalability of your Data Classification deployment.

**IMPORTANT!** For the large-size and extra-large environments, it is strongly recommended to configure a cluster of several NDC Servers and apply DQS mode to these clustered servers. See Configuring NDC Servers Cluster and Load Balancing with DQS Mode for details.

2.2.3.1. Example: Mid-Size Data Environment

This example provides the results of different data processing modes testing for a mid-size environment. The following infrastructure components were deployed as VMware VMs: Netwrix Data Classification server, database server, data source (file server).

2.2.3.1.1. Configuration

**Netwrix Data Classification Server**

<table>
<thead>
<tr>
<th>Specification</th>
<th>Settings</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform</td>
<td>VMware ESXi 6.0</td>
<td></td>
</tr>
<tr>
<td>Hardware:</td>
<td><strong>CPU</strong>: Intel Xeon E5-2683 v4 , 2.10 GHz</td>
<td>Using faster processor increases data processing performance.</td>
</tr>
<tr>
<td></td>
<td><strong>Logical Processors</strong>: 32 (2 sockets, 16 cores per socket)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Memory</strong>: 256 GB</td>
<td></td>
</tr>
<tr>
<td>Virtual machine</td>
<td><strong>CPU</strong>: 8 vCPU</td>
<td></td>
</tr>
<tr>
<td>configuration</td>
<td><strong>Memory</strong>: 32 Gb</td>
<td></td>
</tr>
</tbody>
</table>
## Deployment Specification

<table>
<thead>
<tr>
<th>Specification</th>
<th>Settings</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hard disk</strong></td>
<td><em>SSD storage; thin provisioning enabled</em></td>
<td></td>
</tr>
<tr>
<td><strong>Guest OS</strong></td>
<td>Windows Server 2012 R2 (64-bit)</td>
<td></td>
</tr>
<tr>
<td><strong>Software</strong></td>
<td>Netwrix Data Classification server with <em>Distributed Query Server Mode</em></td>
<td>Used 4 server instances with <em>Distributed Query Server Mode</em> configuration. See &quot;Distributed Query Server Mode&quot; for details.</td>
</tr>
</tbody>
</table>

### Database Server

<table>
<thead>
<tr>
<th>Specification</th>
<th>Settings</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform</td>
<td>VMware ESXi 6.0</td>
<td></td>
</tr>
<tr>
<td><strong>Hardware:</strong></td>
<td><em>CPU:</em> Intel Xeon E5-2660 v4 , 2.00 GHz</td>
<td>Using faster processor increases data processing performance.</td>
</tr>
<tr>
<td></td>
<td><em>Logical Processors:</em> 56 (2 sockets, 14 cores per socket)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Memory:</em> 512 GB</td>
<td></td>
</tr>
<tr>
<td><strong>Virtual machine configuration</strong></td>
<td><em>CPU:</em> 8 vCPU</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Memory:</em> 128 Gb</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Hard disk:</em> SSD storage; thin provisioning enabled</td>
<td></td>
</tr>
<tr>
<td><strong>Guest OS</strong></td>
<td>Windows Server 2012 R2 (64-bit)</td>
<td></td>
</tr>
<tr>
<td><strong>Software</strong></td>
<td>Microsoft SQL Server 2016 SP2 Enterprise Edition</td>
<td></td>
</tr>
</tbody>
</table>

### Data Source (File Server)

<table>
<thead>
<tr>
<th>Specification</th>
<th>Settings</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform</td>
<td>VMware ESXi 6.7</td>
<td></td>
</tr>
<tr>
<td><strong>Hardware:</strong></td>
<td><em>CPU:</em> Intel Xeon E5-2620 v4 , 2.10 GHz</td>
<td>Using faster processor increases data processing performance.</td>
</tr>
</tbody>
</table>
2.2.3.1.2. Data Set

The file server with the following data set was used as a content source:

- Number of files: 1,000,000+
- Number of folders: 65,000
- File types: PDF, DOCX, HTML, RTF, TXT
- Average file size: 500 K - 1 MB
- Total data set size: 1.8 TB

2.2.3.1.3. Data Processing

Data processing was launched for the file server with 1,000,000+ objects (files and folders) in each mode: No Index, Keyword, Compound Term. It was set up to use all predefined taxonomies, no OCR.

Automated workflow was configured as follows:

- Workflow condition: a file gets classified with any taxonomy (in addition to the Size, Type and Language standard taxonomies).
- Workflow rule: such file is migrated to the dedicated location.

Data processing results for all modes are listed below.

<table>
<thead>
<tr>
<th>Specification</th>
<th>Settings</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Logical Processors</strong></td>
<td>32 (2 sockets, 8 cores per socket)</td>
<td></td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>128 GB</td>
<td></td>
</tr>
<tr>
<td><strong>Virtual machine configuration</strong></td>
<td>CPU: 8 vCPU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Memory: 32 Gb</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hard disk: SSD storage; thin provisioning enabled</td>
<td></td>
</tr>
<tr>
<td><strong>Guest OS</strong></td>
<td>Windows Server 2019 Standard (64-bit)</td>
<td></td>
</tr>
</tbody>
</table>
2. Deployment

### Files processed per minute (average)

<table>
<thead>
<tr>
<th>Specification</th>
<th>No Index</th>
<th>Keyword</th>
<th>Compound Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Files processed per minute (average)</td>
<td>558</td>
<td>514</td>
<td>479</td>
</tr>
<tr>
<td>Files with workflow condition triggered (i.e. at least 1 taxonomy applied)</td>
<td>135584</td>
<td>154888</td>
<td>152524</td>
</tr>
<tr>
<td>NDC SQL database size (MDF)*</td>
<td>9.5 GB</td>
<td>9.8 GB</td>
<td>7.2 GB</td>
</tr>
<tr>
<td>Index size</td>
<td>55 GB</td>
<td>176 GB</td>
<td>321 GB</td>
</tr>
</tbody>
</table>

* — For overall space estimations, secondary files should also be considered, so please contact your database administrator.

### 2.2.3.2. Example: Large-Size Environment

This example provides the results of different data processing modes testing for a larger data environment. The following infrastructure components were deployed as VMware VMs: Netwrix Data Classification server, database server, data source (file server).

#### 2.2.3.2.1. Configuration

**Netwrix Data Classification Server**

<table>
<thead>
<tr>
<th>Specification</th>
<th>Settings</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform</td>
<td>VMware ESXi 6.7</td>
<td>Using faster processor increases data processing performance.</td>
</tr>
<tr>
<td>Hardware:</td>
<td><strong>CPU</strong>: Intel Xeon E5-2660 v4 , 2.00 GHz</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Logical Processors</strong>: 56 (2 sockets, 14 cores per socket)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Memory</strong>: 256 GB</td>
<td></td>
</tr>
<tr>
<td>Virtual machine hardware</td>
<td><strong>CPU</strong>: 8 vCPU</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Memory</strong>: 32 Gb</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Hard disk</strong>: SSD storage; thin provisioning enabled</td>
<td></td>
</tr>
<tr>
<td>Guest OS</td>
<td>Windows Server 2012 R2 (64-bit)</td>
<td></td>
</tr>
</tbody>
</table>
## Deployment Specification

### Database Server

<table>
<thead>
<tr>
<th>Specification</th>
<th>Settings</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Software</strong></td>
<td>Netwrix Data Classification server with <em>Distributed Query Server Mode</em> configuration.</td>
<td>Used 4 server instances with <em>Distributed Query Server Mode</em> configuration. See &quot;Distributed Query Server Mode&quot; for details.</td>
</tr>
</tbody>
</table>

#### Data Source (File Server)

<table>
<thead>
<tr>
<th>Specification</th>
<th>Settings</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Platform</strong></td>
<td>VMware ESXi 6.7</td>
<td>Using faster processor increases data processing performance.</td>
</tr>
<tr>
<td><strong>Hardware</strong></td>
<td><strong>CPU</strong>: Intel Xeon E5-2620 v4, 2.10 GHz</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Logical Processors</strong>: 32 (2 sockets, 8 cores per socket)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Memory</strong>: 128 GB</td>
<td></td>
</tr>
<tr>
<td><strong>Virtual machine hardware</strong></td>
<td><strong>CPU</strong>: 8 vCPU</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Memory</strong>: 128 GB</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Hard disk</strong>: SSD storage; thin provisioning enabled</td>
<td></td>
</tr>
<tr>
<td><strong>Guest OS</strong></td>
<td>Windows Server 2012 R2 (64-bit)</td>
<td></td>
</tr>
<tr>
<td><strong>Software</strong></td>
<td>Microsoft SQL Server 2016 SP2 Enterprise Edition</td>
<td></td>
</tr>
</tbody>
</table>
### 2.2.3.2.2. Data Set

The file server with the following data set was used as a content source:

- Number of files: 32,000,000+
- Number of folder: 2,000,000+
- File types: PDF, DOCX, HTML, RTF, TXT
- Average file size: 500K - 1MB
- Total data set size: 57 TB

### 2.2.3.2.3. Data Processing

Data processing was launched for the file server with 34,000,000+ objects (files and folders) in **Keyword** mode. It was set up to use all predefined taxonomies, no OCR.

Automated workflow was configured as follows:

- Workflow condition: a file gets classified with any taxonomy (in addition to the Size, Type and Language standard taxonomies).
- Workflow rule: such file is migrated to the dedicated location.

Data processing results are listed below.

<table>
<thead>
<tr>
<th>Specification</th>
<th>Settings</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Virtual machine hardware</strong></td>
<td><strong>CPU:</strong> 8 vCPU</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Memory:</strong> 32 Gb</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Hard disk:</strong> SSD storage; thin provisioning enabled</td>
<td></td>
</tr>
<tr>
<td><strong>Guest OS</strong></td>
<td><strong>Windows Server 2019 Standard (64-bit)</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Keyword</strong></th>
<th><strong>Processing time</strong></th>
<th>62 days 19 hrs 47 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Files processed per minute (average)</td>
<td></td>
<td>365</td>
</tr>
<tr>
<td>Files with workflow condition triggered (i.e. at least 1 taxonomy applied)</td>
<td></td>
<td>4752724</td>
</tr>
<tr>
<td>NDC SQL database size (MDF)*</td>
<td></td>
<td>190GB</td>
</tr>
<tr>
<td>Index size</td>
<td></td>
<td>4 TB</td>
</tr>
</tbody>
</table>
* — For overall space estimations, secondary files should also be considered, so please contact your database administrator.
2.3. Requirements to Install Netwrix Data Classification

This section contains the hardware and software requirements and other prerequisites needed to deploy Netwrix Data Classification.

- Hardware Requirements
- Software Requirements
- Accounts and Required Permissions
2.3.1. Hardware Requirements

Review the hardware requirements for the computer where Netwrix Data Classification will be installed.

You can deploy Netwrix Data Classification on a virtual machine running Microsoft Windows guest OS on the corresponding virtualization platform, in particular:

- VMware vSphere
- Microsoft Hyper-V
- Nutanix AHV

Note that Netwrix Data Classification supports only Windows OS versions listed in the Software Requirements section.

2.3.1.1. Netwrix Data Classification Server

The requirements in this section apply to a single Netwrix Data Classification server.

To deploy a server cluster, make sure all planned cluster nodes meet the requirements listed below. Consider deploying 1 Netwrix Data Classification Server per approx. 1,000,000 objects to process.

See Deployment Planning and Configuring NDC Servers Cluster and Load Balancing with DQS Mode for more information.

<table>
<thead>
<tr>
<th>Hardware Component</th>
<th>Minimum Requirements</th>
<th>Recommended</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>Any modern.</td>
<td>Any multi-core</td>
</tr>
<tr>
<td></td>
<td>Consider that greater CPU frequency and number of cores improve overall performance of Netwrix Data Classification Server.</td>
<td></td>
</tr>
<tr>
<td>RAM</td>
<td>8 GB</td>
<td>16 GB</td>
</tr>
</tbody>
</table>

2.3.1.2. SQL Server

Review the hardware requirements for the computer where Netwrix Data Classification SQL Database will be deployed.
## 2. Deployment

### 2.3.1.3. Network Access

<table>
<thead>
<tr>
<th>Specification</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network access</td>
<td>Ensure that your Netwrix Data Classification servers are available over the network on a HTTP compliant port from all machines where the client interface (management console) will run.</td>
</tr>
</tbody>
</table>

### 2.3.1.4. Configuring NDC Servers Cluster and Load Balancing with DQS Mode

The **Distributed Query Server (DQS)** mode allows you to balance the load between multiple Netwrix Data Classification Servers (NDC Servers) while data collection, indexing and classification. This approach is strongly recommended if you need to process large data volumes, for example:

- **File Servers**—Up to 32 m objects per cluster of 4 servers.
- **SharePoint**—Up to 8 m objects per cluster of 4 servers.

To apply **Distributed Query Server** mode, you need to arrange your NDC Servers in a 'cluster' for load distribution, as described below. Each clustered NDC Server will store its own set of .CSE files — that is, **NDC Index** will be a distributed index. To assemble and combine data required for the search results, each NDC Server will automatically communicate with the other clustered servers.

**NOTE:** All NDC Servers in the cluster will share a single NDC SQL database.
This functionality is implemented through the *QueryServer* application installed together with NDC Server.

### 2.3.1.4.1. Applying DQS Mode

DQS mode can be configured via the administrative web console.

If you want to implement DQS configuration for your NDC deployment, consider the following:

- This action cannot easily be undone, so before applying the DQS mode, take a full backup of your NDC deployment. Also, read the related documentation sections thoroughly before you start.
- Make sure all servers you plan to add to the DQS cluster have proper network connection and are visible to each other across the network. Adjust firewall settings if necessary.
- Initially, all existing documents will be ‘allocated’ to the first server in the ‘cluster’ and then re-distributed across all configured servers.

To be able to configure the DQS mode, current account requires a **Superuser** role.

**To arrange NDC Servers cluster and apply DQS mode**

1. Install and configure the first Netwrix Data Classification Server as described in the [Install Netwrix Data Classification](#) section.
2. Open administrative web console.
3. Navigate to **Config → Utilities → DQS**.
4. Select **Enable DQS**.

**NOTE:** Once the DQS mode is enabled, you cannot roll back your configuration. Netwrix strongly recommends to ensure that you have taken a full backup of your environment. If ready, confirm the DOS enablement operation when prompted.

5. On the **DQS** tab, click **Add** to add servers you prepared, one by one.

Complete the following fields:
2. Deployment

Setting | Value
--- | ---
Server | Provide the NDC Server name or IP address (name format is case-insensitive).
QS Path | Path to the solution component responsible for DQS mode, residing on the server being added. Filled in automatically; leave the default value.
Active | Select to enable clustering for the instance being added.
Alternate Server | Netwrix recommends using default values.
Alternate QS Path | Netwrix recommends using default values.

6. Click **Save** to close the dialog.

7. Prepare to install other Netwrix Data Classification Server instances, assuming each server requires a dedicated machine. Make sure they meet the [Hardware Requirements](#) and general [Software Requirements](#).

8. On each server, follow the installation steps as described in the [Install Netwrix Data Classification](#) section until **SQL Database** configuration.

9. On the **SQL Database** step, provide the name of the SQL Server instance that hosts **NDC SQL database** you configured for the first NDC Server.

   **NOTE:** Ignore the confirmation dialog on the existing schema in the selected SQL database.

10. Complete the installation.

11. Repeat steps 2 - 6 for every NDC Server, then review the list of servers to make sure the new server was included.
12. If you were configuring the DQS mode for the existing NDC deployment, you will be prompted to re-collect data from the data sources—in order to re-distribute the content index across all NDC Servers in the cluster.

**NOTE:** To force re-distribution when necessary, you can use the **Re-Collect** command available after clicking **Run Cleaner** button on the **Config > Settings > Collector** tab.

To review system health and check your configuration, use the product dashboards. See [Review Dashboards](#) for more information.
### 2.3.2. Software Requirements

The table below lists the software requirements for the Netwrix Data Classification installation:

<table>
<thead>
<tr>
<th>Component</th>
<th>Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating system</td>
<td>Windows 2012 R2 and above Server Operating System Software.</td>
</tr>
<tr>
<td>Windows Features</td>
<td>Web Server Role (IIS)</td>
</tr>
<tr>
<td>Common HTTP Features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Default Document</td>
</tr>
<tr>
<td></td>
<td>• HTTP Errors</td>
</tr>
<tr>
<td></td>
<td>• Static Content</td>
</tr>
<tr>
<td></td>
<td>• HTTP Redirection</td>
</tr>
<tr>
<td>Security</td>
<td>• Windows Authentication</td>
</tr>
<tr>
<td></td>
<td>• Anonymous Authentication</td>
</tr>
<tr>
<td></td>
<td><strong>NOTE:</strong> The Anonymous Authentication element is included in the default installation of IIS 7. Make sure you use IIS 7 and above.</td>
</tr>
<tr>
<td>Application</td>
<td>• ISAPI Extensions</td>
</tr>
<tr>
<td>Development</td>
<td>• ISAPI Filters</td>
</tr>
<tr>
<td>Other features</td>
<td>• .NET Framework 4.7.2</td>
</tr>
<tr>
<td></td>
<td>• ASP.NET</td>
</tr>
<tr>
<td>SQL Server</td>
<td>• SQL Server 2008 R2 Standard Edition (or later).</td>
</tr>
<tr>
<td></td>
<td>• SQL Server 2016 SP2 recommended (for better performance).</td>
</tr>
<tr>
<td></td>
<td><strong>NOTE:</strong> For large environments, SQL Server Enterprise edition may be needed; see needed. See <a href="#">Deployment Planning</a>.</td>
</tr>
</tbody>
</table>
## Component Requirements

<table>
<thead>
<tr>
<th>Component</th>
<th>Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft IFilters</td>
<td>• <a href="#">Microsoft Office 2010 Filter Packs</a> and above, 64-x edition.</td>
</tr>
</tbody>
</table>

### Other software

**Antivirus**

Netwrix recommends adding Netwrix Data Classification Index files to the list of exclusions (white list) of any installed antivirus. These files have `.CSE` extension.
2.3.3. Accounts and Required Permissions

Netwrix Data Classification uses the following accounts:

<table>
<thead>
<tr>
<th>Account</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service Account</td>
<td>This account is specified during the product setup. Windows domain account</td>
</tr>
<tr>
<td></td>
<td>that you plan to use as a service account will need the following:</td>
</tr>
<tr>
<td></td>
<td>• <strong>Local Administrator</strong> rights on the server where Netwrix Data</td>
</tr>
<tr>
<td></td>
<td>Classification will be installed.</td>
</tr>
<tr>
<td></td>
<td>• Permissions to run the <strong>Windows Services</strong> and <strong>IIS Application</strong> pool.</td>
</tr>
<tr>
<td></td>
<td>After installation, this account will be automatically granted the <strong>Logon</strong></td>
</tr>
<tr>
<td></td>
<td>as a <strong>service privilege</strong> on the Netwrix Data Classification server.</td>
</tr>
<tr>
<td></td>
<td>• SQL Server <strong>DBO</strong> permissions to the NDC SQL database (if using Windows</td>
</tr>
<tr>
<td></td>
<td>Authentication to access SQL Server).</td>
</tr>
</tbody>
</table>

**NOTE:** Optionally, you can use local account instead of domain account.

| Crawl content       | Ensure the availability of accounts with sufficient permissions to         |
|                    | access your content sources:                                             |
|                    | • SharePoint, SharePoint Online site collection— **Site Collection**        |
|                    | **Administrator** role.                                                   |
|                    | • Exchange mailboxes:                                                     |
|                    | 1. **ApplicationImpersonation**—allows the crawling account to impersonate |
|                    | each of the mailboxes / users configured for collection.                   |
|                    | 2. **Mailbox Search**—allows the crawling account to enumerate mailboxes, |
|                    | i.e. automatic discovery of mailboxes.                                    |

See **Configure Microsoft Exchange for Crawling and Classification** for detailed information on configuring these permissions.

• Outlook Mail Archive (PST file)— **Read** permission.

• File System (SMB, NFS) — **Read** permission for the folders and files you need to crawl.
## 2. Deployment

<table>
<thead>
<tr>
<th>Account</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• G Suite and Google Drive — service account needs permissions to read data in the individual and shared Drives on behalf of users using the Google Drive API.</td>
</tr>
<tr>
<td></td>
<td>See <a href="#">Configure G Suite for Crawling</a> for detailed information.</td>
</tr>
<tr>
<td></td>
<td>• Database— <strong>Read</strong> permission for the database schema and data.</td>
</tr>
</tbody>
</table>

| Apply tagging    | To use tagging, i.e. to write classification attributes back to the content file, service account will need the appropriate **Modify** permissions on the content source. |
2.4. Configure NDC Database

Netwrix Data Classification uses Microsoft SQL Server database as data storage. During installation, you have been prompted to create a dedicated **NDC SQL database** on your SQL Server instance. Upon installation completion, you need to configure it as shown below for the product to function properly. You can create the database manually prior to the product installation—Using **SQL Server Management Studio** or **Transact-SQL**. Refer to the following Microsoft article for detailed instructions on how to create a new database: [Create a Database](#).

**NOTE:** For performance purposes, Netwrix strongly recommends to separate NDC and SQL Server machine.

For certain product features, SQL Server Standard or Enterprise edition is required.

To configure NDC database

**NOTE:** The account used to create the NDC SQL database must be granted the **dbcreator** server-level role.

1. On the computer where SQL Server instance with the **NDC SQL database** resides, navigate to **Start** → **All Programs** → **Microsoft SQL Server** → **SQL Server Management Studio**.
2. Connect to the server.
3. Locate the **NDC_Database**, right-click it and select **Properties**.
4. Select the **Files** page and set the **Initial Size (MB)** parameter for PRIMARY file group to **512 MB**.
5. Click **Expand** next to **PRIMARY** file group and set **Autogrowth / Maxsize** as follows:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>File Growth</td>
<td>• Recommended—128 MB.</td>
</tr>
<tr>
<td></td>
<td>• Large environment—512 MB.</td>
</tr>
<tr>
<td>Maximum File Size</td>
<td>Select <strong>Unlimited</strong>.</td>
</tr>
</tbody>
</table>

6. Go to **Options** page and make sure that the **Recovery model** parameter is set to "Simple".

**NOTE:** Netwrix recommends that you do not change the recovery model to avoid log files growth.
2.5. Install Netwrix Data Classification

1. Run `Netwrix_Data_Classification.exe`.
2. Review minimum system requirements and then read the License Agreement. Click Next.
3. Follow the instructions of the setup wizard. When prompted, accept the license agreement.
4. On the Product Settings step, specify path to install Netwrix Data Classification. For example, `C:\Program Files\NDC`.
5. On the Configuration step, specify the directory where Index files reside. For example, `C:\Program Files\NDC\Index`.
6. On the SQL Database step, provide SQL Server database connection details.

Complete the following fields:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server Name</td>
<td>Provide the name of the SQL Server instance that hosts your NDC SQL database. For example, &quot;WORKSTATIONSQL\SQLSERVER&quot;.</td>
</tr>
<tr>
<td>Authentication Method</td>
<td>Select Windows or SQL Server authentication method.</td>
</tr>
<tr>
<td>Username</td>
<td>Specify the account name.</td>
</tr>
<tr>
<td>Password</td>
<td>Provide your password.</td>
</tr>
<tr>
<td>Database Name</td>
<td>Enter the name of the SQL Server database. Netwrix recommends using NDC_ database name.</td>
</tr>
</tbody>
</table>

7. On the Licensing step, add license. You can add license as follows:
   - Click the Import button and browse for your license file
   - OR
   - Open your license file with any text editor, e.g., Notepad and paste the license text to the License field.
8. On the Administration Web Application step, review default IIS configuration.
9. On the Services step, configure Netwrix Data Classification services:
   - Select all services to be installed.
   - File System Path—Use default path or provide a custom one to store Netwrix Data Classification's Services files. For example, `C:\Program Files\NDC Services`. 
• Provide user name and password for the product services service account.

**NOTE:** This account is granted the Logon as a service privilege automatically on the computer where NDC is going to be installed.

• Select additional service options, if necessary.

10. On the Pre-Installation Tasks and Checks step, review your configuration and select **Install**.

11. When the installation completes, open a web browser and navigate to the following URL: http://localhost/conceptQS where localhost is the name or IP address of the computer where Netwrix Data Classification is installed. For example, http://workstationndc/conceptQS.

### 2.6. Upgrade to the Latest Version

Netwrix recommends that you upgrade from the older versions of Netwrix Data Classification to the latest version available in order to take advantage of the new features.

**NOTE:** Seamless upgrade to Netwrix Data Classification 5.5.2 is supported for versions 5.5.1. If you need to upgrade from an earlier version, please perform staged upgrade, e.g., 5.5.0 → 5.5.1 → 5.5.2.

#### 2.6.1. Take Preparatory Steps

Before you start the upgrade, it is strongly recommended to take the following steps:

1. **IMPORTANT!** Make sure you have .NET Framework 4.7.2 installed on the computer where Netwrix Data Classification resides. If not, download it from Microsoft website: [Download .NET Framework 4.7.2](https://dotnet.microsoft.com/download/dotnet-framework/4.7.2).

2. Back up NDC SQL database. For that:
   - a. Start Microsoft SQL Server Management Studio and connect to SQL Server instance hosting this database.
   - b. In Object Explorer, right-click the database and select Tasks → **Back Up**.
   - c. Wait for the process to complete.

3. Back up the Index files.

4. Finally, close administrative web console.

#### 2.6.2. Considerations and Limitations

During the seamless upgrade from previous versions, Netwrix Data Classification preserves its configuration, so you will be able to classify your data right after finishing the upgrade. However, there are some considerations you should examine - they refer to product operation after upgrading from version 5.5.1:
After the upgrade, you will have to update taxonomies manually. For that:

a. In administrative web console, navigate to Taxonomies → Global Settings.

b. Click Update in the right corner next to each taxonomy.

After the upgrade, indexing mode will be set to Compound Term mode. Refer to the following Netwrix knowledge base article for instructions on how to modify default Index Processing Mode:

2.7. Configuring NDC Servers Cluster and Load Balancing with DQS Mode

The Distributed Query Server (DQS) mode allows you to balance the load between multiple Netwrix Data Classification Servers (NDC Servers) while data collection, indexing and classification. This approach is strongly recommended if you need to process large data volumes, for example:

- **File Servers**—Up to 32 m objects per cluster of 4 servers.
- **SharePoint**—Up to 8 m objects per cluster of 4 servers.

To apply Distributed Query Server mode, you need to arrange your NDC Servers in a ‘cluster’ for load distribution, as described below. Each clustered NDC Server will store its own set of .CSE files — that is, NDC Index will be a distributed index. To assemble and combine data required for the search results, each NDC Server will automatically communicate with the other clustered servers.

**NOTE:** All NDC Servers in the cluster will share a single NDC SQL database.

This functionality is implemented through the QueryServer application installed together with NDC Server.

2.7.1. Applying DQS Mode

DQS mode can be configured via the administrative web console.
If you want to implement DQS configuration for your NDC deployment, consider the following:

- This action cannot easily be undone, so before applying the DQS mode, take a full backup of your NDC deployment. Also, read the related documentation sections thoroughly before you start.
- Make sure all servers you plan to add to the DQS cluster have proper network connection and are visible to each other across the network. Adjust firewall settings if necessary.
- Initially, all existing documents will be 'allocated' to the first server in the 'cluster' and then redistributed across all configured servers.

To be able to configure the DQS mode, current account requires a Superuser role.

**To arrange NDC Servers cluster and apply DQS mode**

1. Install and configure the first Netwrix Data Classification Server as described in the Install Netwrix Data Classification section.
2. Open administrative web console.
3. Navigate to Config → Utilities → DQS.
4. Select Enable DQS.

**NOTE:** Once the DQS mode is enabled, you cannot roll back your configuration. Netwrix strongly recommends to ensure that you have taken a full backup of your environment. If ready, confirm the DOS enablement operation when prompted.
5. On the DQS tab, click Add to add servers you prepared, one by one.

Complete the following fields:
6. Click **Save** to close the dialog.

7. Prepare to install other Netwrix Data Classification Server instances, assuming each server requires a dedicated machine. Make sure they meet the [Hardware Requirements](#) and general [Software Requirements](#).

8. On each server, follow the installation steps as described in the [Install Netwrix Data Classification](#) section until **SQL Database** configuration.

9. On the **SQL Database** step, provide the name of the SQL Server instance that hosts **NDC SQL database** you configured for the first NDC Server.

   **NOTE:** Ignore the confirmation dialog on the existing schema in the selected SQL database.

10. Complete the installation.

11. Repeat steps 2 - 6 for every NDC Server, then review the list of servers to make sure the new server was included.

---

**Table: Setting and Value**

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Server</strong></td>
<td>Provide the NDC Server name or IP address (name format is case-insensitive).</td>
</tr>
<tr>
<td><strong>QS Path</strong></td>
<td>Path to the solution component responsible for DQS mode, residing on the server being added. Filled in automatically; leave the default value.</td>
</tr>
<tr>
<td><strong>Active</strong></td>
<td>Select to enable clustering for the instance being added.</td>
</tr>
<tr>
<td><strong>Alternate Server</strong></td>
<td>Netwrix recommends using default values.</td>
</tr>
<tr>
<td><strong>Alternate QS Path</strong></td>
<td>Netwrix recommends using default values.</td>
</tr>
</tbody>
</table>

---

**Diagram:**
12. If you were configuring the DQS mode for the existing NDC deployment, you will be prompted to re-collect data from the data sources — in order to re-distribute the content index across all NDC Servers in the cluster.

**NOTE:** To force re-distribution when necessary, you can use the Re-Collect command available after clicking Run Cleaner button on the Config > Settings > Collector tab.

To review system health and check your configuration, use the product dashboards. See Review Dashboards for more information.
2.8. Configure IT Infrastructure

Successful crawling requires certain configuration of your IT infrastructure, which may include enabling Windows services, etc.

Review the following for additional information:

- Configure Microsoft Exchange for Crawling and Classification
- Configure NFS File Share for Crawling
- Configure G Suite for Crawling
2.8.1. Configure Microsoft Exchange for Crawling and Classification

When crawling an Exchange Server, it is necessary to configure sufficient permissions to allow the crawling account to impersonate the mailboxes that you wish to crawl. This requires the setup of two permissions:

- **ApplicationImpersonation**—Allows the crawling account to impersonate each of the mailboxes / users configured for collection
- **Mailbox Search**—Allows the crawling account to enumerate mailboxes (automatic discovery of mailboxes)

Review the following for additional information:

- To configure using Office 365 Exchange Admin Portal
- To configure using Exchange 2010 or later (on-premise)
- To configure match rules

**To configure using Office 365 Exchange Admin Portal**

1. Login to the [Office 365 Exchange Admin Portal](https://admin.microsoft.com)
2. Go to Permissions, then under admin roles click the '+' symbol to add a new role and enter the Name and Description 'NetwrixCrawlerImpersonation'.
3. Click the '+' symbol under Roles, select ApplicationImpersonation and Mailbox Search roles.
4. Click **add** → and then **OK**.

5. Click the ‘+’ symbol under **Members**: and select your Admin User.

6. Click **add** → then **OK**.

**To configure using Exchange 2010 or later (on-premise)**

1. Login to one of the **Exchange** servers (RDP)

2. Open a **Powershell** window

3. Run the following commands (replacing **ADMINUSERNAME** with the username of your crawling account):

   ```powershell
   New-ManagementRoleAssignment -Name "NetwrixCrawlerImpersonation" -Role "ApplicationImpersonation" -User ADMINUSERNAME
   New-ManagementRoleAssignment -Name "NetwrixCrawlerSearch" -Role "Mailbox Search" -User ADMINUSERNAME
   ```

If crawling **Microsoft Office 365 for Small Business** or many hosted Exchange systems, then it is not possible to setup **Application Impersonation**.

**To configure match rules**
The **Match Rules** are an important configuration option, defining which mailboxes will be crawled as part of an Exchange Server source. Here are some example match rules that may be required:

1. `.*@netwrix.com`— Identifies the domain (netwrix.com) within the mailbox name, restricts crawling to a specific set of mailboxes

2. `.*`— Identifies any mailbox, ensuring that all mailboxes will be crawled.
2.8.2. Configure NFS File Share for Crawling

To process NFS file shares, it is necessary to enable specific Windows features. The steps to enable these features differ depending on operating system of the computer where Netwrix Data Classification is installed.

Consider the following:

- NFS File shares are only supported for the machines running Windows Server 2012 and later (server OS) or Windows 10 and later (workstation OS)
- Changes made to files (including adding new files) will not be automatically detected until the source is re-indexed—Netwrix recommends setting the re-index period for NFS file shares to 1 day.

To configure Windows Server 2012 or later

1. On the Windows desktop, start Server Manager.
2. On the Manage menu, click Add Roles and Features.
3. Progress to the Features step.
4. Ensure that Client for NFS option enabled.
5. Complete the wizard.

To configure Windows 10

1. Navigate to Control Panel and select Programs.
2. Select Turn Windows features on or off.
3. Expand Services for NFS and enable the Client for NFS option.
4. Click OK.

After configuring your NFS share, you will be able to add the Folder content source, as described in the File System section.

NOTE: Do not specify username and password while adding data source.

2.8.3. Configure G Suite for Crawling

Netwrix Data Classification for Google Drive uses the OAuth 2.0 protocol to authenticate to your G Suite domain. You will need to create a service account and authorize it to access data in individual and shared Drives on behalf of users using the Google Drive API. Do the following:

In Google API Console:
1. Create a new project
2. Select Application type
3. Create a new service account
4. Create a service account key (JSON, save a copy for the data source configuration)
5. Enable G Suite domain-wide delegation for the service account (write down the Client ID)
6. Enable Google Drive API

**In G Suite Admin Console:**
1. Authorize service account to access the Google Drive API

**To configure G Suite for crawling**

**IMPORTANT!** Google administrative interfaces tend to change over time, so refer to the following guide for up-to-date instructions on creating OAuth 2.0 service accounts: [Using OAuth 2.0 for Server to Server Applications](#).

Review the following for additional information:

<table>
<thead>
<tr>
<th>To...</th>
<th>Do...</th>
</tr>
</thead>
</table>
| Create a new project | 1. Navigate to https://console.developers.google.com (Google API Console) while logged in as a G-Suite administrator within the domain to be crawled (if the user is not added within the correct domain then the correct data will not be identified).  
2. Create a new project. |
| Select Application type | 1. Once a new project has been created, navigate to APIs & Services → OAuth consent screen.  
2. Set User type to "Internal".  
3. Provide the name for new application.  
4. Click Save. |
| Create a new service account | 1. In Google API console, navigate to IAM & Admin → Service Accounts.  
2. Create service account as described in Google official article.  
3. On the Grant this service account access to project (optional) step, do not select any roles.  
4. On the Grant users access to this service account (optional) step, do not grant any user access. Click Done. |
## 2. Deployment

<table>
<thead>
<tr>
<th>To...</th>
<th>Do...</th>
</tr>
</thead>
</table>
| Create a service account key | 1. On the **Service accounts** page, select the account you want to create a key for.  
2. Click ☰ icon under **Actions** and select **Create key**.  
3. In the **Create private key for <Service account name>** dialog, select **JSON** format, and download the file to a known location as it will be required later.  

**NOTE:** Your new public / private keypair is generated and downloaded to your machine; it serves as the only copy of this key. You are responsible for storing it securely. If you lose this keypair, you will need to generate a new one. |
| Delegate domain-wide authority to the service account | 1. On the **Service accounts** page, select your service account and click **Edit**.  
2. Click the **Show Domain-Wide Delegation** link and tick the **Enable G Suite Domain-wide Delegation** checkbox.  
3. Click **Save**.  
4. Once completed, review the "**Domain wide delegation**" column for this account and make sure that it enabled.  
5. Click the **View Client ID** link.  
6. Copy your Client ID, you will need it later. |
| Enable Google Drive API | 1. In **Google API** console, navigate to the **API Dashboard** and select **Enable APIs and Services** (if APIs have not previously been enabled).  
2. Search for Google Drive API and click **Enable** (or **Manage**).  
3. Switch to **G Suite Admin Console**.  
4. Navigate to **Security → Advanced Settings → Manage API Client Access** within the Google admin portal.  
5. Set the client name to the **Client ID** you copied on the previous step.  
6. Set the API scope to “**https://www.googleapis.com/auth/drive**” and select **Authorize**. |
2.9. Initial Product Configuration

The **Product Configuration Wizard** allows you quickly configure basic Netwrix Data Classification settings such as processing mode, taxonomies, etc.

In your web browser, navigate to the following URL: http://hostname/conceptQS where **hostname** is the name or IP address of the computer where Netwrix Data Classification is installed and perform initial configuration steps.

On the **Instance** step, provide the unique name for your Netwrix Data Classification instance. For example, "Production".

Click **Next** to proceed. See also:

- **Select Processing Mode**
- **Processing Settings**
- **Add Taxonomy**
- **Review Your Configuration**

2.9.1. Select Processing Mode

At this step of the wizard, select processing (indexing) mode for your environment.
For starter and evaluation purpose, select **Keyword** mode.

### 2.9.2. Processing Settings

On the **Processing Settings** step, review options for data processing and classification. For test and evaluation purposes, Netwrix recommends use default values.

Proceed with adding taxonomies.

### 2.9.3. Add Taxonomy

On this step, you are prompted to load predefined taxonomies.
Click the search bar and select one or several taxonomies you want to add. See Built-in Taxonomies Overview for the full list of built-in taxonomies supported by Netwrix Data Classification.

2.9.4. Review Your Configuration

On this step, review your configuration. Once you complete the wizard, you can:

- Add a Source
- Add a Taxonomy
- Take the Product Tour
- Get Help
3. Security (Users)

The Users administration area provides a web based console for creating and managing users who are authorized to use the various administrative functions. It also provides a central mechanism to manage passwords used by the core services to crawl content, as well as the ability to restrict access to the available APIs.

By default no users are defined and usage of the administrative functions built into the QS is unrestricted. You must add at least one user in order to restrict the access to the QS administrative functions.

The QS supports the following types of authentication mechanisms: Windows, ADFS, Azure AD and Forms.

Review the following for additional information:

- Secure Netwrix Data Classification
- User Management
- Password Manager
- Web Service Security

3.1. Secure Netwrix Data Classification

The steps described within this guide can be used to review the security of your Netwrix Data Classification deployment and apply any changes you feel necessary to secure the administration of, and access to, the Classification interfaces.

To configure Administration Console Access

By default, post installation, all users will be considered Superusers with access to all areas of the product. To begin the process of securing the product please follow the below steps:

1. Access the Administration Console
2. Select Users from the top navigation bar
3. Select Add
4. Your username will be pre-filled and must first be added to ensure that you do not lose access to the system.

5. You can now add other users / groups as required - either as Superusers, or with access to specific areas / functions

**Superusers** have access to all areas / functions within the product but may not see all search results (if they have been filtered based on security in the source system such as SharePoint).

Optionally, you can also consider using a federated authentication mechanism, such as Azure AD.

**To configure Microsoft SQL Server Security**

Netwrix Data Classification supports several options to mitigate risk against the content stored in the backend SQL Server database:

- **Connection Encryption**—Protects your data as it moves between the core products and the SQL Server database. To enable this mode, do the following:
  1. Open `conceptConfig` in each of the configured application locations, typically:
     - `C:\inetpub\wwwroot\conceptQS\bin`
     - `C:\Program Files\Concept Searching\Services\ConceptCollectorService`
     - `C:\Program Files\Concept Searching\Services\conceptIndexer`
  2. Check the Encrypt Connection box as well as the Trust Server Certificate box if you do not have a valid certificate loaded for SQL Server.
  3. Click Save.

- **Transparent Data Encryption (TDE)**—Protects your data at rest within SQL Server. Netwrix Data Classification supports the use of TDE, it should of course be noted that this does incur a performance overhead. TDE should be managed and configured by your database administrator(s).

**To secure Search Index (CSE File)**

The **CSE** file index contains the full text search behind the Classification engine. There are two key groupings to this engine:

- **Text.cse**—Stores the raw text of each document in a compressed and obfuscated format.

- **All other files**—Stores the compound term processing search index, identifying which documents should be returned for a given query

**Text.cse** can be optionally encrypted, this utilises AES/SHA256 to further improve the security of the full text at rest. You can enable this mode by:

1. Access the **Administration Console**;
2. Select **Config** from the top navigation bar;
3. Enable the **Encrypt Text (Text.cse)** option (under advanced settings - select the screwdriver spanner to show);

4. Select **Save**.

The remaining files cannot be reverse engineered to retrieve the full document text - however, do contain the weightings and terms within the text. We recommend restricting access to all files at the file system level as well as considering file system encryption.

**To review web service endpoints**

There are several web service endpoints which provide access to various levels of information within Netwrix Data Classification. If you are exposing the administration interface to the internet then you may wish to fully restrict access to these endpoint(s) via your firewall or IIS Configuration (potentially removing all external access).

The following paths should be considered as part of this process:

- /_api/*
- *.asmx
- *.svc

It should be noted that when using Netwrix Data Classification for SharePoint Online certain endpoints are required, each of these endpoints are located within the folder "/ConceptClassifierApp/".

### 3.2. User Management

This section contains information on how to configure user authentication mechanisms, their permissions and manage existing users. Review the following for additional information:

- **Authentication Mechanisms**
- **Adding/Removing Users**
- **Permission Management**
- **Super Users**

**Authentication Mechanisms**

On first install the QS will be configured for Windows authentication. To setup the QS to use an ADFS server please follow the "Installation and Configuration" guide using the section "ADFS". To use forms based authentication please disable all other authentication methods in IIS other than: Anonymous and Forms:
To utilize Azure AD simply create the client application then add two new appSettings to the web.config found in the QS directory:

- `<add key="ida:AzureClientId" value="NewAzureADClientID (GUID)" />
- `<add key="ida:AzureAuthority" value="AzureADAuthorityValue such as: https://login.windows.net/mytenant.onmicrosoft.com" />

The Netwrix Data Classification REST APIs also support Bearer based authentication, to enable this mode please add one further appSetting entry into the web.config file:

- `<key="ida:AzureTenant" value="Tenant Name such as: netwrix.com" />

In certain sections of the QS settings are split between Basic and Advanced. Users wishing to always see Advanced options can enable this by:

- Selecting their username from the footer of the application
- Clicking User Preferences
- Ticking Always Show Advanced Settings
- Clicking Save

**Adding/Removing Users**

More users can be added at any time from the default Users screen, as well as allowing for users to be removed.
Additional Windows users can be validated using Integrated Windows Authentication. Additional non-Windows users can only be added if the Non-Windows Authentication mode is enabled.

If the only user defined is a Super User and that user is deleted then all security is removed and usage of the QS administrative functions reverts to unrestricted.

User accounts granted access to the REST APIs will still be restricted by their specific user permissions. A Superuser with REST API access will be able to run any API method, any normal user will be restricted by the same rules that govern the UI. Further API samples and documentation can be found at: /conceptQS/_api

**Permission Management**

In order to allocate granular permissions to a user (non-Super Users), simply select their username from the main grid.

Each tab contains a top level checkbox (“Allow Access”) which defines whether or not a user has access to each of the top level administrative areas.

When an area is enabled there are typically more granular permissions that can be enabled, such as:

- Within the **Taxonomies** area it is also possible to assign permissions at a specific Term Set or Term branch level. A full user permission summary (for all Term/Set level permissions) can be viewed by selecting the **View Taxonomy Permissions** button (shown below).

- Within the **Sources** area it is possible to restrict a user’s access to specific source groups, as shown below.

```
<table>
<thead>
<tr>
<th>Users</th>
<th>Access Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Allowed Source Groups</td>
</tr>
<tr>
<td></td>
<td>Source Management</td>
</tr>
<tr>
<td></td>
<td>Page Management</td>
</tr>
<tr>
<td></td>
<td>Actions</td>
</tr>
<tr>
<td></td>
<td>All Source Groups</td>
</tr>
<tr>
<td></td>
<td>Add</td>
</tr>
<tr>
<td></td>
<td>Edit Advanced Config</td>
</tr>
<tr>
<td></td>
<td>Delete</td>
</tr>
<tr>
<td></td>
<td>Delete</td>
</tr>
<tr>
<td></td>
<td>Source Management</td>
</tr>
<tr>
<td></td>
<td>Page Management</td>
</tr>
<tr>
<td></td>
<td>Actions</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Taxonomy Permissions Summary:
Super Users

Super Users always have access to all Query Server administrative functions.

Non-Super Users must have their access rights specifically configured and all rights are disabled by default. See Permission Management for details about configuring the access rights for non-Super Users.

Regardless of the authentication mode selected the usage of the QS administrative functions will continue to be unrestricted until at least one user is added. The first user must be a Super User. If Windows or ADFS Authentication are being used then the first user will default to the currently logged in user, although this can be changed if required.

If Non-Windows Authentication is enabled then additional information must be entered to define the non-Windows user.

3.3. Password Manager

Password manager can be used to automatically schedule password changes, for service accounts that are being used to access external systems. This is particularly useful when there are business policies in place to change passwords on a rolling basis.
To amend the passwords for a username record first select **Passwords** from the main display. Then either click **Edit** on a particular password row, or, click **Add Password** to add a new password for the account. It is not possible to have overlapping date ranges for the defined passwords, nor is it possible to remove all passwords from a user record.

### 3.4. Web Service Security

**Web Service Security** can be used to restrict external access to the Netwrix Data Classification APIs, we recommend when using this functionality that you list the Netwrix Data Classification service account under the **Allow Only Listed** records. When **Block All** is selected certain functionality within Netwrix Data Classification will be impacted (if there is API use).

Certain methods must be individually enabled for security reasons, such as **GetSourceItemContent** which allows you to retrieve the binary content of a crawled item.

There are three modes available:

- **Allow All**—No restrictions, all users have access to the APIs
- **Block All**—No API use supported
- **Allow Only Listed**—Blocks all API use except for those users (or groups) listed

Each mode is assigned to a specific grouping of service methods, you can see which API functions are affected by clicking the “View Methods” link and edit the security mode by clicking the **Edit** link.
4. Content Sources

A content source is a repository of data presented within Netwrix Data Classification to be crawled and classified. Each source has an individual configuration and, where appropriate, credentials.

For adding and managing content systems, use the Sources area of the Netwrix Data Classification management console. You can manage the individual content sources or organize them into source groups, which are used as logical containers.

You can configure the unlimited number of sources to work with.

**IMPORTANT!** To access the Sources area, users require sufficient rights. See the User Management section for more information.

See next:

- Add a Content Source
- Modify Source Settings
- Manage Sources and Control Data Processing
- Source Groups

4.1. Add a Content Source

To start processing your data, you need to add a corresponding content source to the Netwrix Data Classification scope.

*To add a content source*

1. In administrative web console, navigate to Sources → General and click Add to launch the Add source wizard.
2. Select the source you need and configure its settings. See detailed instructions for the sources:

- **Database** (Microsoft SQL Server or Oracle database)
- **Exchange Server** or **Exchange Mailbox**
- **File System** (includes Folder and File)
- **Add Google Drive Source**
- **Outlook Mail Archive**
- **SharePoint** or **SharePoint Online**

All your content sources will be listed in the **Sources** section.

**NOTE:** When adding a source or managing source configuration, the most commonly used source settings are displayed by default. However, some source types have additional configuration options that can be displayed by clicking the **Advanced Settings** ("wrench" icon). You can allow these advanced settings to be always shown to authorized users. See **Security (Users)** for more information.

### 4.1.1. Database

It is also possible to index a wide variety of other sources, including:

- Microsoft SQL Server
- Oracle Databases
- PostGres Databases
- EMC Documentum DMS
- Interwoven Worksite DMS
- Hummingbird DMS
Content must either be configured / crawled using the configured service accounts (IIS Application Pool User, Windows Services) or by using specific connection details. For PostGres connections the username/password must be specified.

Once connected it is possible to create an intelligent content mapping, crawling certain fields as unstructured index text, and other fields as mapped metadata. For more information please see the Manage SQL section.

If you wish to make other configuration changes before collection of the source occurs ensure you tick the checkbox "Pause source on creation".

Complete the following fields:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connection Type</td>
<td>Select your connection type: MS SQL, MySQL, Oracle, or PostgreSQL.</td>
</tr>
<tr>
<td>Server</td>
<td>Specify the server name of the database system to be crawled (&quot;.&quot; can be used to indicate the local server).</td>
</tr>
<tr>
<td>Database Name</td>
<td>Specify the database that will be crawled. It is possible to configure multiple databases from the same server.</td>
</tr>
<tr>
<td>Authentication Method</td>
<td>Select your authentication method.</td>
</tr>
<tr>
<td>Source Group</td>
<td>If you want to add database to a source group, select existing, or create a new one. See Grouping Sources for more information.</td>
</tr>
<tr>
<td>Pause source on creation</td>
<td>Select to make other configuration changes before collection of the source occurs.</td>
</tr>
</tbody>
</table>

When the connection configuration has been completed you will be redirected to the Source Configuration, this allows you to define how the database will be crawled. It is possible to crawl either specific tables, or crawl custom queries (defined select statements, which may use JOIN statements across multiple tables).

### 4.1.2. Exchange Mailbox

The Exchange source configuration screen allows you to enable the crawling and classification of content stored in a single Exchange mailbox.

Complete the following fields:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email Address / Password</td>
<td>The Email Address / Password combination can either be for the mailbox desired for crawling, or, can be an administrator account that has been assigned the right of Impersonation.</td>
</tr>
</tbody>
</table>
### 4. Content Sources

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>By default crawling will attempt to locate the correct Exchange URL by using the <strong>Exchange Auto Discover</strong> functionality. Where this is not available the API URL should be specified. This is typically in the format: <code>https://servername/EWS/Exchange.asmx</code>.</td>
</tr>
<tr>
<td>Mailbox</td>
<td>When using impersonation the <strong>Mailbox</strong> should be specified as the mailbox to be crawled, for example: <strong>Email Address</strong> set to <a href="mailto:administrator@cs.com">administrator@cs.com</a>, and <strong>Mailbox</strong> set to <a href="mailto:test@cs.com">test@cs.com</a>.</td>
</tr>
<tr>
<td>Crawl Range</td>
<td>Defines how the data should be accessed from Exchange, selecting a date range will crawl a static set of data, whereas using the <strong>Since</strong> mode will periodically re-crawl from the <strong>Since</strong> date, taking into account the last crawl date for each artifact.</td>
</tr>
<tr>
<td>Re-Index Period</td>
<td>Specifies how often the source should be checked for changes. The number specifies the period in days.</td>
</tr>
<tr>
<td>Build Search index</td>
<td>Specifies whether the mail items should be available from the Netwrix Data Classification index – when disabled classification will occur as normal, but items will not be retrievable from search.</td>
</tr>
<tr>
<td>Document Type</td>
<td>Used to specify a value which can be used to restrict queries when utilizing the Netwrix Data Classification search index.</td>
</tr>
<tr>
<td>Pause source on creation</td>
<td>Select if you want to make other configuration changes before collection of the source occurs.</td>
</tr>
</tbody>
</table>

### 4.1.3. Exchange Server

The Exchange Server source configuration screen allows you to enable the crawling and classification of multiple Exchange mailboxes from the same Exchange server.

It is also possible to provide a filter expression to ensure that certain Mailboxes are included and others excluded as required.

Complete the following fields:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email Address / Password</td>
<td>The Email Address / Password combination must be an administrator account that has been assigned the right of <strong>Impersonation</strong> as well as the <strong>Discovery Management</strong> role.</td>
</tr>
</tbody>
</table>

By default crawling will attempt to locate the correct Exchange URL by
Option | Description
--- | ---
 | using the **Exchange Auto Discover** functionality. Where this is not available the API URL should be specified. This is typically in the format: https://servername/EWS/Exchange.asmx.

| Crawl Range | Defines how the data should be accessed from Exchange, selecting a date range will crawl a static set of data, whereas using the **Since** mode will periodically re-crawl from the **Since** date, taking into account the last crawl date for each artifact.

| Detection Period | Specifies how often the source should be checked for changes. The number specifies the period in days.

### 4.1.4. File System

Use the Source configuration screen to set up the crawling and classification operations for content stored in your file server. There are two options to configure a content source: as individual files or as folders. Select, respectively, **File** or **Folder** in the Add content source wizard.

#### 4.1.4.1. Add Folder source

Use **Folder** to add the following content sources:

- Windows folders
- SMB (CIFS) shares
- NFS shares
**IMPORTANT!** To add NFS share, make sure you have configured it for crawling as described in [Configure NFS File Share for Crawling](#).

By default, configuration window displays basic configuration settings only. To configure advanced settings, click the "wrench" icon in the bottom left corner.

**NOTE:** To configure advanced settings, your user account will need advanced privileges. See [Security](#) (Users) for more information.

Complete the following fields:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic settings</strong></td>
<td></td>
</tr>
<tr>
<td>Folder</td>
<td>Enter the UNC path of the root folder where collection is to start.</td>
</tr>
<tr>
<td>Depth Limit</td>
<td>Specify how many levels the indexing should process. Possible options:</td>
</tr>
<tr>
<td></td>
<td>• <strong>Exclude Subfolders</strong></td>
</tr>
<tr>
<td></td>
<td>• <strong>All Subfolders</strong> (default setting)</td>
</tr>
<tr>
<td></td>
<td>• <strong>Limit Subfolders</strong> - if selected, specify the required subfolders depth (from 2 to 99)</td>
</tr>
<tr>
<td>Write classifications</td>
<td>Select if you wish to write classifications directly into the document properties, i.e. use tagging. This applies to DOC/DOCX/XLS/XLSX/PPT/PPTX/PDF.</td>
</tr>
<tr>
<td></td>
<td>See also <a href="#">Set up filters and tagging for File System</a>.</td>
</tr>
<tr>
<td>Source Group</td>
<td>Default value recommended.</td>
</tr>
<tr>
<td>Pause source on creation</td>
<td>Select if you want to make other configuration changes before collection of the source occurs.</td>
</tr>
<tr>
<td><strong>Advanced settings</strong></td>
<td></td>
</tr>
<tr>
<td>Username</td>
<td>Specify the account used to process the folder.</td>
</tr>
<tr>
<td>Password</td>
<td>Provide a password for the account specified above.</td>
</tr>
<tr>
<td>Text Patterns</td>
<td>See <a href="#">Text Handling</a> for more information.</td>
</tr>
<tr>
<td>Date Filter</td>
<td>Use this calendar control to instruct the program to only crawl the content that has been modified since the specified date.</td>
</tr>
<tr>
<td></td>
<td>This can be useful for targeting data that is current - in situations where there is a huge volume of content (assuming that the most recent content has the highest risk).</td>
</tr>
</tbody>
</table>
4. Content Sources

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anonymous Access Allowed</td>
<td>Select this option to disable security filtering for the content source. If cleared, the indexing processes will collect Windows Access Control Lists (ACLs) for the files, and search results will be filtered based upon the end user's Windows identity.</td>
</tr>
<tr>
<td>Duplicate Detection Enabled</td>
<td>Select to exclude duplicates (i.e. documents that contain the same text content) from the index.</td>
</tr>
<tr>
<td>Re-Index Period</td>
<td>Specifies how often the source should be checked for changes. Netwrix recommends using default values. Default is 7 days.</td>
</tr>
<tr>
<td>Priority</td>
<td>Netwrix recommends using default values.</td>
</tr>
<tr>
<td>Document Type</td>
<td>Specify a value that will be used to restrict queries when utilising the search index.</td>
</tr>
</tbody>
</table>

When finished, click **Save**.

**4.1.4.2. Add Files source**

Use the **File** section to crawl individual files.

By default, configuration window displays basic configuration settings only. To configure advanced settings, click the "wrench" icon in the bottom left corner.

**NOTE:** To configure advanced settings, your user account will need advanced privileges. See **Security (Users)** for more information.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>File Source</td>
<td>Select how you wish to provide the file location:</td>
</tr>
<tr>
<td></td>
<td>○ <strong>File</strong> - enter file path</td>
</tr>
</tbody>
</table>
## 4. Content Sources

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>o <strong>Browse</strong> - browse for the file you need</td>
<td><strong>Advanced settings</strong></td>
</tr>
<tr>
<td>Source Group</td>
<td>Default value recommended.</td>
</tr>
<tr>
<td><strong>Username</strong></td>
<td>Specify the account used to process the file.</td>
</tr>
<tr>
<td><strong>Password</strong></td>
<td>Provide a password for the account specified above.</td>
</tr>
<tr>
<td><strong>Anonymous Access Allowed</strong></td>
<td>Select this option to disable security filtering for the content source.</td>
</tr>
<tr>
<td></td>
<td>If cleared, the indexing processes will collect Windows Access Control Lists (ACLs) for the files, and search results will be filtered based upon the end user’s Windows identity.</td>
</tr>
<tr>
<td><strong>Upload</strong></td>
<td>If selected, the file will be uploaded into the NDC SQL database. This will allow the program to present the file to users even if they do not have access to the original file location.</td>
</tr>
<tr>
<td><strong>Text Patterns</strong></td>
<td>See <a href="#">Text Handling</a> for more information.</td>
</tr>
<tr>
<td><strong>Max Collector Retries</strong></td>
<td>Specify how many retries are attempted before automatically removing items from the index when incremental collection indicates that the file has been deleted. Default is <strong>3</strong> retries.</td>
</tr>
<tr>
<td><strong>Re-Index Period</strong></td>
<td>Specifies how often the source should be checked for changes. Netwrix recommends using default values. Default is <strong>7 days</strong>.</td>
</tr>
<tr>
<td><strong>Priority</strong></td>
<td>Netwrix recommends using default values.</td>
</tr>
<tr>
<td><strong>Document Type</strong></td>
<td>Specify a value that will be used to restrict queries when utilising the search index.</td>
</tr>
</tbody>
</table>

### 4.1.5. Add Google Drive Source

The Google Drive source configuration screen allows you to enable the crawling and classification of content stored in both G-Suite repositories and Google Drive personal accounts.

**IMPORTANT!** Make sure you created App for GDrive crawling prior to start adding the source. See [Configure G Suite for Crawling](#) for more information.
Complete the following fields:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Drive Type</strong></td>
<td>Select <em>Business</em>.</td>
</tr>
<tr>
<td><strong>User Email(s)</strong></td>
<td>When adding a G-Suite source, enter the email address of the user's drive that you wish to crawl (via impersonation).</td>
</tr>
<tr>
<td><strong>Crawl Shared Items</strong></td>
<td>Select to crawl all files shared with the specified user in addition to any team drives shared with the user.</td>
</tr>
<tr>
<td><strong>Crawl Shared Items</strong></td>
<td>Select to enable crawling of any types of documents shared with the specified user.</td>
</tr>
<tr>
<td><strong>JSON Import</strong></td>
<td>Drag the JSON connection file you downloaded while creating Google service account in the form.</td>
</tr>
<tr>
<td><strong>Project ID</strong></td>
<td>Open the JSON connection file and copy file contents to <strong>Project ID</strong> field.</td>
</tr>
<tr>
<td><strong>Write Classifications</strong></td>
<td>Leave this checkbox empty.</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Source Group</td>
<td>Netwrix recommends creating a dedicated source group for Google Drive.</td>
</tr>
<tr>
<td>Pause source on creation</td>
<td>Select if you want to make other configuration changes before collection of the source occurs.</td>
</tr>
</tbody>
</table>

4.1.6. Outlook Mail Archive

The Outlook Mail Archive source configuration screen allows you to enable the crawling and classification of content stored in PST files:

NOTE: If you wish to make other configuration changes before collection of the source occurs ensure you tick the checkbox **Pause source on creation**.

Multiple mailboxes can be added at one time via the "+" button. Collection will process all folders / emails / attachments within the mailbox - associating the attachment text with the respective email.

Folders / Items can be excluded from processing via the Exchange Exclusions management screen.

4.1.7. SharePoint

The SharePoint section allows for one or more site collections to be queued for processing that share the same set of crawling credentials.

The following versions of SharePoint are supported: 2010, 2013, 2016, 2019 and SharePoint Online.

If you wish to make other configuration changes before collection of the source occurs ensure you tick the checkbox **Pause source on creation**.
Complete the following fields:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SharePoint URL</td>
<td>The root of the site collections to be added, by clicking the “(Multiple URLs)” link you can add multiple SharePoint Site Collections to be crawled against the same credentials.</td>
</tr>
<tr>
<td>Username</td>
<td>Enter username in the following formats: DOMAIN\USERNAME and USERNAME@DOMAIN.</td>
</tr>
<tr>
<td>Write classifications to SharePoint</td>
<td>Enables synchronization of classifications back to the SharePoint managed metadata fields. The written classifications will be subject to the classification configuration for the site collection.</td>
</tr>
<tr>
<td>Re-Index Period</td>
<td>Specifies how often the source should be checked for changes. The number specifies the period in days.</td>
</tr>
<tr>
<td>Document Type</td>
<td>Specify a value which can be used to restrict queries when utilizing the Netwrix Data Classification search index.</td>
</tr>
</tbody>
</table>

### 4.1.8. SharePoint Online

Office 365 customers can configure the collector service to automatically detect and queue their employees OneDrive (Personal Sites) hosted in Office 365. An account with Tenant administration rights must be supplied, and the frequency of the detection of new OneDrive sites must be set. It is also possible to provide a filter expression to ensure that certain OneDrive paths are included and others excluded as required.

Optionally, it is also possible to set up the resources necessary to ensure Netwrix Data Classification Classifier is enabled and configured on the detected OneDrive sites. Templating allows an administrator to pre-configure classification configurations for site collections. For more information please review the associated templating guide.
4.2. Narrow Data Collection Scope

Inclusions and exclusions provide a granular way of limiting collection scope to a specific set of documents within a content source.

This functionality is currently supported for the following source types:

- Exchange
- File System
- Google Drive

See next:
Set up exclusions and tagging for Exchange
Set up filters and tagging for File System
Set up exclusions and tagging for Google Drive

4.3. Use Tagging (optional)

Tagging in Netwrix Data Classification means writing classification attributes back to the content files. Tagging enables external systems (that is, not directly integrated with Netwrix Data Classification) to leverage the automatically generated classifications for a variety of business purposes, for example:

- Enriching the search experience
- Driving the application of DLP/Security labelling
Tagging is designed to work as natively as possible with each source type. Therefore, each integration varies in the way that classifications can be written, with some overlaps.

Typically, to use tagging, you need to take the following steps:

1. Ensure that an appropriate license has been loaded to enable document tagging. For that, go to Config → Licensing → Licensing Summary.
2. Ensure that the credentials you plan to use for accessing the source system have been granted the appropriate Modify permissions.
3. Ensure that tagging has been enabled for the source objects— for that, select the Write Classifications option in the source settings.
4. Configure the source-specific settings to map the classifications results back to the source properties, as described in the related section.

NOTE: If you are unsure of the correct source specific settings to use, then we recommend initially working with some sandbox data.

You can Pause source processing while you are configuring the correct settings to ensure that no tagging will occur with partial/incorrect configuration settings.

See also:
- Set up granular processing and tagging for Database
- Set up exclusions and tagging for Exchange
- Set up filters and tagging for File System
- Set up exclusions and tagging for Google Drive
- Set up processing options for SharePoint

### 4.4. Manage Sources and Control Data Processing

The following commands are available on the General tab of the Sources section:

- **Delete**—Removes the source from processing. Its content will not appear in the search results in due course.

NOTE: This does not delete content from the external system

- **Re-Collect**—Queues the source for re-processing. Crawled items will be deleted, and the entire source re-crawled
- **Re-Index**—Queues a source or item to be re-indexed, with a check for changes; if changes are found, the item will be re-indexed

- **Re-Classify**—Queues a source or item to be re-classified against the latest configured classification rules

**NOTE:** See [Index Maintenance](#) for more information on these operations.

- **Pause**—Temporarily pauses source content processing

- **Resume**—Resumes a source from a temporary pause

- **Add To Group**—Adds a source to a logical container (Source Group), either an existing or a newly created one.

Besides, in the source list on the **General** tab you can do the following for selected source:

- **View Results**

- **Edit** the source details by clicking on the "gear" icon

- **View source-specific statistics** by clicking on the "chart" icon

- **View detailed information** by clicking on the “i” icon

- **Navigate to the source** by clicking on the “link” icon

**NOTE:** When adding a source or managing source configuration, the most commonly used source settings are displayed by default. However, some source types have additional configuration options that can be displayed by clicking the **Advanced Settings** ("wrench" icon). You can allow these advanced settings to be always shown to authorized users. See [Security (Users)](#) for more information.

### 4.4.1. Modify Source Settings

To edit configuration settings for the certain source, select the source and go to the corresponding tab, e.g. **Box** or **SharePoint**. Then you can, in particular, specify **Write configuration** (i.e. “tagging”) settings and apply source-specific parameters. See [Use Tagging (optional)](#) for more information.

See also:
4. Content Sources

- Set up granular processing and tagging for Database
- Set up exclusions and tagging for Exchange
- Set up filters and tagging for File System
- Set up exclusions and tagging for Google Drive
- Set up processing options for SharePoint

4.4.2. Set up granular processing and tagging for Database

This section contains information on how to configure granular classification and crawling of your databases. For example, you can specify which tables / views / queries will be crawled, or set up table configuration.

Also, you can use Write Configuration options to configure "tagging". See the following:

- Source Configuration
- Primary Key Query
- Content Query
- Table Configuration
- Write Configuration

**Source Configuration**

The Source Configuration screen allows you to define which tables / views / queries will be crawled. The following options are available:

- Add Source—Add a new SQL database connection
- Edit Connection—Amend the connection details of the currently selected source
- Add Query—Add a custom method for crawling content (custom SELECT statements), Templates are provided for Hummingbird, Worksite and Documentum.

You can access the Source Configuration screen by selecting the multi-cog (Advanced Configuration) icon from the sources grid.

Selecting Edit for one of the tables / queries on the list will redirect you to the entity level configuration, which identifies how content will be mapped into the core index.
Selecting the **Add Query** option will present a popup allowing you to select a unique name for the query, as well as the queries to be used for crawling:

**Primary Key Query**

The primary key query should return a set of values that uniquely identify each row to be crawled, in the event that **JOINs** are used you should **JOIN** from the largest dataset to the smallest, to ensure that each row is unique.

**Example:**

```
SELECT PageID FROM Pages
```

**NOTE:** Stored procedures are currently not supported.

**Content Query**

- Select a **Table** for crawling.
- Ensure **Enabled** is set to **true**.
- Use the **Add Query** option to add unique names and queries for crawling.

### Table Example:

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Enabled</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>dbo.SecurityAudit</code></td>
<td>Table</td>
<td>X</td>
<td>Edit</td>
</tr>
<tr>
<td><code>dbo.ACLGroupMembership</code></td>
<td>Table</td>
<td>X</td>
<td>Edit</td>
</tr>
<tr>
<td><code>dbo.ACLs</code></td>
<td>Table</td>
<td>X</td>
<td>Edit</td>
</tr>
<tr>
<td><code>dbo.ACUserMembership</code></td>
<td>Table</td>
<td>X</td>
<td>Edit</td>
</tr>
<tr>
<td><code>dbo.AnalyticsQueue</code></td>
<td>Table</td>
<td>X</td>
<td>Edit</td>
</tr>
<tr>
<td><code>dbo.AnalyticsQueue</code></td>
<td>Table</td>
<td>X</td>
<td>Edit</td>
</tr>
<tr>
<td><code>dbo.Attachments</code></td>
<td>Table</td>
<td>X</td>
<td>Edit</td>
</tr>
<tr>
<td><code>dbo.AttachmentsExcluded</code></td>
<td>Table</td>
<td>X</td>
<td>Edit</td>
</tr>
<tr>
<td><code>dbo.AutoClassificationChanges</code></td>
<td>Table</td>
<td>X</td>
<td>Edit</td>
</tr>
<tr>
<td><code>dbo.Backups</code></td>
<td>Table</td>
<td>X</td>
<td>Edit</td>
</tr>
</tbody>
</table>

**NOTE:** Stored procedures are currently not supported.
The content query must return all fields to be indexed/classified on, as well as the fields included in the primary key query.

Example: `SELECT * FROM Pages`

**NOTE:** Stored procedures are currently not supported

Adding the query will take you to the custom query configuration. Here you can update the primary key query and the content query, all other configuration options are described in the Table Configuration section:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Include</td>
<td>When checked the table/entity will be enabled in the collection schema.</td>
</tr>
<tr>
<td>Upload Content</td>
<td>When checked the Content fields will be uploaded into the SQL database.</td>
</tr>
<tr>
<td></td>
<td>Uploaded content can be retrieved after collection by passing the PageId for</td>
</tr>
<tr>
<td></td>
<td>the record to the QS API call &quot;GetDownload&quot;.</td>
</tr>
<tr>
<td>PK - Primary Key</td>
<td>Please select the fields which uniquely identify the row to be crawled, in</td>
</tr>
<tr>
<td></td>
<td>the event that multiple rows are returned by the Primary Key, the query will</td>
</tr>
<tr>
<td></td>
<td>be aborted. Custom queries will not require the primary key to be defined,</td>
</tr>
<tr>
<td></td>
<td>this will be set automatically from the primary key query.</td>
</tr>
<tr>
<td>Content</td>
<td>Identifies the fields that will be crawled as searchable text in the core</td>
</tr>
<tr>
<td></td>
<td>search index. Multiple fields can be mapped to Content, each will be appended</td>
</tr>
<tr>
<td></td>
<td>with a line break.</td>
</tr>
<tr>
<td></td>
<td>It is also possible to configure a single binary field type that contains a</td>
</tr>
<tr>
<td></td>
<td>document, the collection process will load the binary and attempt to convert</td>
</tr>
<tr>
<td></td>
<td>and extract text from the document. When this functionality is used we</td>
</tr>
<tr>
<td></td>
<td>recommend setting the <strong>ContentFilename</strong> or <strong>ContentType</strong> index mapping</td>
</tr>
<tr>
<td></td>
<td>to aid the process of text extraction.</td>
</tr>
</tbody>
</table>
### Option | Description
--- | ---
Metadata | Identifies the fields that will be mapped as metadata values.

Index Mappings | Index mappings identifies mappings between the entities fields and the internal core database. Each row also contains an information icon identifying its purpose within the crawling process.

Modified Filter (Incremental Crawls) | This should be set to a field that defines when a row has changed (the modified date for the row). When set the collection process will automatically filter the re-indexing process to rows that have a modified date that is larger than the last crawl time.

Re-Index Period | This value is the number of days/hours/minutes that will pass between Re-Indexing. The Re-Indexing process involves querying the table(s) to find new and changed records.

---

**Write Configuration**

The SQL write configuration allows you to update a specific column per taxonomy within the source repository with the associated classifications for a record.

### Option | Description
--- | ---
Table Name | Specify the name of the table to be updated (in most cases this will be the
4. Content Sources

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column Name</td>
<td>Specify the name of the column to be updated (text/varchar column).</td>
</tr>
</tbody>
</table>
| Update Filter   | Update filters are the method used to restrict the update at the target destination. If multiple filters are configured then they all must be true. Filters should be created in the format: ColumnName=@Parameter, where @Parameter is a correctly configured metadata value from the source table/query. The specified values will result in a query in the following format:

```
UPDATE TABLENAME SET COLUMNNAME=@Classifications WHERE FILTERS
```

Format          | Specify the delimiters/construction of the value to be written into the SQL database.                                                  |
4.4.3. Set up exclusions and tagging for Exchange

When indexing emails / folders from Exchange the list of locations that will be ignored is defined by the `CollectionExclusions` list. The definitions in this list may be viewed and modified via the `Exclusions` form:

Any item with a name that matches one of these patterns will be ignored. Wildcards may be used anywhere in the pattern definition, with:

- The asterisk character (*) matching any sequence of characters
- The question mark character (?) matching any single character

4.4.4. Set up filters and tagging for File System

This section contains information on how to include or exclude files or folders from being crawled, and how to configure writing classification attributes back to the content files (i.e. "tagging").

4.4.4.1. Configure Inclusions

You can define the list of file locations that should be included when indexing files.

File inclusions are based on file extensions – all inclusions are prepended with "*". Using no wildcard indicators will include a specific extension (i.e. ".PDF").

Ending with a wildcard indicator will match all extensions which start with the inclusion (e.g. ".DOC*" will match DOC, DOCX and DOCM files).

Do the following:

1. In the management console, click **Sources → Box**, then in the left pane click **Files Included**.
2. Select the necessary extensions to be used as including filter when processing files.
3. To modify an extension (for example, add a wildcard), click **Edit**. To add a new one, click **Add**.

4.4.4.2. **Configure Exclusions**

1. In the management console, click **Sources → File**, then in the left pane click **Files Excluded**.

2. In the **Details** window specify the objects (files or folders) to exclude:

   To exclude a certain file, enter its full path. For example: `C:\Test Folder\Test Document.docx`

   Wildcards can be used anywhere in the exclusion pattern definition as follows:
   - The asterisk character (*) matching any sequence of characters
   - The question mark character (?) matching any single character

   For example:
   - `*/Restricted Folder/*` will exclude specific folder in any File source
NOTE: Exclusions are case-insensitive.

3. Optionally, enter a test path to verify the settings and click **Test**.

4. Finally, click **Save** and close the window.

### 4.4.4.3. Configure Tagging

You can instruct the program to write classification attributes back to processed files. This operation is also called "tagging". Tagging is currently supported for the following file types:

- DOC/DOCX
- PPT/PPTX
- XLS/XLSX
- PDF

For Microsoft Office documents, each taxonomy is mapped to an advanced (custom) property in the document's metadata. See [this article](#) for details.

For Adobe PDF documents, each taxonomy is mapped to custom properties in the document's metadata. See [this article](#) for details.

Related content source settings can be configured at a global level (default), or at a source level.

**To configure tagging on a global level**

1. In the management console, click **Sources** → **Box**, then in the left pane click **Write Configuration**.
2. Select the taxonomy you need and click the **Edit** link for it.
3. In the taxonomy properties, enable writing classification attributes (tags) and specify other settings:

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Enabled</strong></td>
<td>Use to enable / disables the writing of classifications for the selected taxonomy.</td>
<td>Cleared by default</td>
</tr>
<tr>
<td><strong>Field Name</strong></td>
<td>Defines the attribute name to be used when persisting the classifications (metadata property name).</td>
<td></td>
</tr>
<tr>
<td><strong>Single Value Field</strong></td>
<td>If selected, this option will cause only the highest scoring classification to be written to the field.</td>
<td></td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td>How the classifications should be formatted.</td>
<td>You can create a custom delimited combination of the labels / GUIDs.</td>
</tr>
</tbody>
</table>
4. Content Sources

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name/ID or Class</td>
<td>Depending on the format, take the term labels, IDs or a combination of both</td>
<td>The corresponding Delimiter must be a string or array type with a maximum length of 3.</td>
</tr>
<tr>
<td>Prefix/ Suffix</td>
<td>Will be appended to the formatted string of classifications.</td>
<td></td>
</tr>
</tbody>
</table>

![Agriculture](image)

To configure tagging on a source level

1. Go to Sources → General, highlight the source you need and click the "pencil" symbol on the right.

2. The list of taxonomy configurations set up globally will be displayed. To apply these global settings, select **Use Global Configuration** check box on top. To configure source-specific settings, clear this check box.

3. Select the taxonomy you need and click **Edit**.

4. In the dialog with taxonomy configuration, select the **Enabled** checkbox and specify the settings described in the table above.
4.4.5. Set up exclusions and tagging for Google Drive

This section contains information on how to write configuration and include items on a Google Drive source. Review the following for additional information:

- Excluded Items
- Write Configuration

**Excluded Items**

When indexing files from Google Drive the list of file locations that will be ignored is defined by the CollectionExclusions list. The definitions in this list may be viewed and modified via the Exclusions form:

Any item with a name that matches one of these patterns will be ignored. Wildcards may be used anywhere in the pattern definition, with:

- The asterisk character (*) matching any sequence of characters
- The question mark character (?) matching any single character

**Write Configuration**
The **Write Configuration** options define how classifications should be written back to the Google Drive:

Settings can be configured at a global level (default), or at a source level by selecting the "pencil" symbol from the default sources screen for a Google Drive source.

Classifications are written back to the document properties in the Google Drive repository. Each taxonomy can be mapped to a single property - though, if it is possible to split classifications across multiple fields if a text limit is hit within the source system.

**NOTE:** Writing classifications to documents in this source will affect additional document metadata such as modified date and/or modified user.

Also note any classifications written to Google Drive are stored in custom properties which are not visible to an end user - they are only accessible via the Google Drive APIs.

### 4.4.6. Set up processing options for SharePoint

This section contains information on how to configure granular classification and crawling of your SharePoint farm (for example, exclude a site from being processed or define custom configuration to your subsite). Review the following for additional information:
Dashboard

The SharePoint dashboard provides the same dashboard display as the main reporting dashboard, with the results filtered to SharePoint types. Classification coverage identifies the percentage of content that has had classifications applied, and the percentage that has not.

Advanced Configuration

The Advanced Configuration screen allows you to define which content within the SharePoint site collection will be crawled. With the following options available:

- Entity Configuration
  - Include / Exclude an entity (subsite/list)
  - Define custom metadata mappings per entity
4. Content Sources

- **Source Defaults**—Specify the default custom metadata mapping for the site collection

- **Configuration Viewer**—Simple XML display to view the raw configuration

Custom metadata mappings allows the user to map specific SharePoint fields to internal indexed fields. There are two types of mapping:

- **Content Field Mappings**—The fields which listed as "Content Fields" will be extracted and indexed when the site collection is spidered/processed by the collector service

- **Special Field Mappings (Including Date fields)**—The purpose of these mappings is to make use of the advanced filtering options available in the core search index (for example: It allows a SharePoint date field to be mapped into the "Last Modified" value - allowing results to be retrieved only if they are in a certain date range)

Mappings operate on a "Defaults" basis. In the absence of a list level configuration the collector service will automatically use the mappings configured at the subsite level or global level if there is no subsite configuration.

You can access the **Source Configuration** screen by selecting the multi-cog (**Advanced Configuration**) icon from the sources grid.

Doing so will load the Advanced Configuration screen for the appropriate source.

**Entity Configuration Grid**

The initially shown grid displays the root level information for the site collection. Subsites are navigable to allow configuring subsites/lists at all levels of the hierarchy.

Each item shows a tick/cross indicating whether the container is configured for crawling, under the **Include** column. Each item also displays an indication for "Has Config?" - which indicates whether custom metadata mappings have been defined.

Lists / Subsites can be excluded on a case by case basis by selecting the appropriate link (**Include / Exclude**) from the grid row.

**NOTE:** Excluding content will not automatically remove content from the index. If content has already been crawled then it should be manually deleted via the QS - or, a re-collect performed. When new content is defined for crawling a re-index operation should be performed.
List Configuration

Selecting Edit for a list / library will present the below interface allowing for a custom configuration to be defined. In the absence of a list level configuration the collector will automatically use the subsite level mapping (on a field by field basis). You can use the dropdown lists/selectors to search for and assign SharePoint fields to the appropriate mappings.

NOTE: Content fields cannot be configured at the subsite level. In the absence of a list level configuration the appropriate source defaults will automatically be used.

Subsite Configuration

Selecting Edit for a subsite will present the below interface allowing for a custom configuration to be defined. In the absence of a subsite level configuration the collector will automatically use the source level mappings (on a field by field basis). You can use the dropdown lists/selectors to search for and assign SharePoint fields to the appropriate mappings.
NOTE: Content fields cannot be configured at the subsite level. In the absence of a list level configuration the appropriate source defaults will automatically be used.

The special field mappings allow you to map any of the available SharePoint fields to some of the internal Concept Searching values for the purposes of search. Mappings of each type will be taken first from the list level settings, before reverting to the subsite level settings and finally the source level settings.

Source Configuration (Defaults)

The Source Configuration tab allows you to configure defaults that will be used in the absence of list / subsite configurations. You can use the dropdown lists/selectors to search for and assign SharePoint fields to the appropriate mappings.

The General configuration options also allow overriding enabling / disabling the write back of classifications as well as specifying a regular reindexing period (more frequent than once per day).

The values configured for each of the default content mappings will be assigned based on the base template of the list (Document Library, Generic List etc).
SharePoint Excluded

When indexing files from SharePoint the list of file locations that will be ignored is defined in the **SharePoint Excluded** list. The definitions in this list may be viewed and modified via the SharePoint Excluded form:

Any item with a name that matches one of these patterns will be ignored. Wildcards may be used anywhere in the pattern definition, with:

- The asterisk character (*) matching any sequence of characters
- The question mark character (?) matching any single character

**Templating**

Templating allows an administrator to pre-configure classification configurations for site collections. For more information please review the associated templating guide.

**4.4.7. Set up processing options for SharePoint Online Tenancy**

Typically SharePoint environments are crawled on a per site collection basis. Sometimes however there is a need to crawl an entire SharePoint Online tenancy. The following guide details the step-by-step instructions in order to configure a whole tenancy for collection.
1. Add SharePoint Online source as described in the SharePoint Online section.

   **NOTE:** If this option is not available within the source type selection then it would suggest that the source type is not currently licensed, please contact support for more details.

2. The Source is configured to the tenancy level, therefore we recommend specifying the URL as the root site collection URL. This is however not a requirement if you do not have a root site collection.

3. Specify an account with tenancy administration rights. Accounts can be specified in either the default AD format \DOMAIN\USERNAME, or in the format of the user’s email address USERNAME@DOMAIN.

4. The Match Rules are an important configuration option, defining which site collections will be crawled. Here are some example match rules that may be required:
   - .*/Personal/.*—Identifying "/personal/" within the URL (as per the below example) - this would be the correct configuration to crawl end-user’s OneDrive site collections (OneDrive for Business)
   - .*— Identifies any site collections, ensuring that all collections will be crawled

5. Define the required Classification Template, as well as the Detection Period which defines how often we will detect new site collections

6. Select Save.

### 4.5. View Results

#### 4.5.1. Data Processing Statistics

Select the source from the list on the Sources - General tab, and click the Reports ("chart") icon to view data processing statistics for that source.

#### 4.5.2. Content Crawling and Classification Results

Click on a source row in the list of sources on the General tab to view the crawled data, including the number of processed documents/URLs (Documents column), the size of the crawled content (Size), status, etc.

To browse the whole structure of the crawled content, click on the items in the list. It is also possible to filter the list by any field.
Each document has an associated status (shown as the ID). Click the numeric ID to read the status description:

- Click the "Info" icon for the document/item to view its **Properties**, including summary, classifications (if any), etc.:
• For content sources that support writing the classifications back to the source system, i.e. "tagging" (e.g. such as writing classifications to SharePoint managed metadata fields):
  
  • a tick will also be displayed if tagging was successful
  
  • a cross displayed if tagging failed

See the related content source description for details.
5. Taxonomies

5.1. What are Taxonomies?

Netwrix Data Classification comes with several built-in taxonomies with hundreds of classification rules out-of-the-box. The taxonomies cover a broad range of sensitive personal, financial, and health-related information. Each taxonomy contains a set of terms. Terms are defined by set of configuration rules (also called clues). See Classification Rules (Clues) for details.

- To create a taxonomy, go to the Taxonomies area of the web-based management console and follow the procedures described in Add a Taxonomy section.

- To manage the taxonomies, follow the procedures described in Manage Taxonomies section.

IMPORTANT! To access the Taxonomies area, users require sufficient rights. See the User Management section for more information.

See also:

- Built-in Taxonomies Overview

- Taxonomy Settings

5.2. Built-in Taxonomies Overview

Netwrix Data Classification comes with eight taxonomies with hundreds of classification rules out-of-the-box.
The four core taxonomies cover a broad range of sensitive personal, financial, and health-related information. The remaining four taxonomies derive from the core set. They are tailored to meet the requirements of specific data protection regulations:

- Personally identifiable information covering GDPR scope.
- Medical records covering HIPAA scope.
- Financial records and payment cards information covering GLBA and PCI DSS scope.

This section contains the full list of built-in taxonomies supported by Netwrix Data Classification.

### 5.2.1. Core Taxonomies

#### Financial Records

- ABA routing numbers
- IBAN/SWIFT codes
- Bank account numbers

#### Personally Identifiable Information (PII)

- Personal information (full name, home address, date of birth) in the following languages:
  - Danish
  - Dutch
  - English
  - French
  - German
  - Greek
  - Icelandic
  - Italian
  - Slovenian
  - Spanish
  - Swedish
- National IDs, passport numbers, driver licenses, taxpayer IDs, etc. for the following countries (coverage varies):
  - Australia
  - Belgium
  - Brazil
- Bulgaria
- Canada
- Croatia
- Cyprus
- Czech Republic
- Denmark
- Estonia
- Finland
- France
- Germany
- Greece
- Hong Kong
- Hungary
- Iceland
- India
- Ireland
- Italy
- Latvia
- Lithuania
- Luxemburg
- Malta
- Netherlands
- Norway
- Poland
- Portugal
- Romania
- Russia
- Singapore
- Slovakia
- Slovenia
- South Africa
5. Taxonomies

- Spain
- Sweden
- United Kingdom
- USA

**Payment Card Industry Data Security Standard (PCI DSS)**

Cardholder data (holder name, card number, expiration and security code) for the major payment systems:

- American Express
- Diners Club
- Discover
- JCB
- Mastercard
- UnionPay
- Visa

**Patient Health Information (PHI)**

Medical forms, treatment records, prescription drugs, decease names/codes, allergies, social and insurance numbers.

### 5.2.2. Derived Taxonomies

**General Data Protection Regulation (GDPR)**

A subset of the PII taxonomy relating to the personal information of EU residents:

- Austria
- Belgium
- Bulgaria
- Croatia
- Czech Republic
- Denmark
- Estonia
- Finland
- France
- Germany
• Greece
• Hungary
• Ireland
• Italy
• Latvia
• Lithuania
• Luxemburg
• Malta
• Netherlands
• Poland
• Portugal
• Romania
• Russia
• Slovakia
• Slovenia
• Spain
• Sweden
• United Kingdom
• Austria
• Belgium
• Bulgaria
• Croatis
• Czech Republic
• Denmark
• Estonia
• Finland
• France
• Germany
• Greece
• Hungary
• Ireland
• Italy
5. Taxonomies

- Latvia
- Lithuania
- Luxemburg
- Malta
- Netherlands
- Poland
- Portugal
- Romania
- Russia
- Slovakia
- Slovenia
- Spain
- Sweden
- United Kingdom

**GDPR Restricted**

Personal data (same as in PII) accompanied by the following special categories of personal information (GDPR Article 9):

- Ethnicity
- Political views
- Religious beliefs

**Gramm-Leach-Bliley Act (GLBA)**

Combines the Financial Records, PCI DSS and PII (US social security numbers) taxonomies.

**Health Insurance Portability and Accountability Act (HIPAA)**

Combines the PHI and PII (US social security numbers) taxonomies.

### 5.3. Taxonomy Settings

This section contains information about taxonomies settings. Review the following for additional information:

- [Taxonomy Settings Levels](#)
- [Labels](#)
5.3.1. Taxonomy Settings Levels

Review the following for additional information:

- Taxonomy/TermSet Level
- Class / Term Level

**Taxonomy/TermSet Level**

When the root node is selected in the treeview (the termset) the Settings tab will display top level taxonomy settings as well as global settings applicable to the Taxonomies area.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Filters</td>
<td>This field allows the taxonomy to be restricted based on a boolean filter (e.g. using the “CSE-FOLDERS” field) or any of the 8 documentid filters. See the associated design guide for more information about the ContentFilter field in the Taxonomies table.</td>
</tr>
<tr>
<td>Max Categories</td>
<td>Sets the maximum number of classes from this taxonomy that will be</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>allocated to each document. To set the Max Categories value across all taxonomies use the Settings tab in Index Manager.</td>
</tr>
<tr>
<td>Default Threshold</td>
<td>Sets the default threshold for newly created terms within the selected taxonomy (does not affect existing terms).</td>
</tr>
<tr>
<td>Create Default Clues</td>
<td>This setting controls the creation of default clues when. If enabled then a default clue is added to all Classes based on the title of the class – or, optionally based on the default metadata clue format.</td>
</tr>
<tr>
<td>Default Clue Score</td>
<td>Sets the default score value for new clues.</td>
</tr>
<tr>
<td>Default Metadata Clue</td>
<td>Specifies the format of a default metadata clue. This can be used to create automatic “self-referential” clues, as well as static assignments based on the term name in the document metadata. “[TermName]” can be utilised for a dynamic lookup of the classes name.</td>
</tr>
<tr>
<td>Count Mode</td>
<td>Sets the display mode for counts in the treeview.</td>
</tr>
<tr>
<td>Show Empty Nodes</td>
<td>Sets the display mode for empty nodes in the treeview.</td>
</tr>
<tr>
<td>Synchronise Termset</td>
<td>Enables/Disables automatic synchronisation through the TermStoreManager tool for the whole Term Set.</td>
</tr>
</tbody>
</table>

**Class / Term Level**

When a child node is selected in the treeview the “Settings” tab will display settings for the selected term:
5. Taxonomies

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available for Tagging</td>
<td>Use to prevent any documents getting classified against a class. This would normally only be set to “No” when a class is being used to boost another class – see Term Boosts for information on terms that use the “Term Boost” type clues.</td>
</tr>
<tr>
<td>Synchronise Term</td>
<td>Enables / Disables automatic synchronisation through the TermStoreManager tool for the term and its children.</td>
</tr>
<tr>
<td>Relevance Threshold</td>
<td>The threshold for each Class defaults to 50 – but can be raised (to reduce the number of documents that get classified) or lowered (to increase the number of documents that get classified).</td>
</tr>
</tbody>
</table>
| Boosts                  | The Weighting Boosts can also be adjusted for each Class. Based on the values above you would expect a 10% score boost if one of its child terms was classified.  
It is possible to set the “Child” boost to 100%, doing so will in effect enable the parent to always be tagged if the child is tagged. An example for this would be a taxonomy containing regions, if a document was tagged as “England” it should also be tagged as “Europe”. |

5.3.2. Labels

This section contains information on how to configure SharePoint and Office 365 labels.
5.3.2.1. SharePoint Labels

SharePoint labels (Alternate Term Labels) are alternate labels configured in SharePoint against the English language. Through the administration interface it is possible to add and remove alternate labels. It is not currently possible to change the default label (this should be achieved by renaming the node via the treeview right click menu).

5.3.2.2. O365 Labels

For a simple automated experience it is possible to assign Office 365 Classification labels to existing Term Set structures within Taxonomy Manager.

At the time of classification the classification process will identify any terms that have both met their threshold and also contain mappings to Office Classification Labels. The engine will then select the highest scoring term, and automatically apply the mapped label to the document in SharePoint (taking into account which labels are available per site collection as well as the setting specified at the term level).

More than one label can be applied to each term to allow for labels to be applied that are only available on a limited set of site collections.

Simply select Add and choose the label you wish to assign from the drop down list:

NOTE: If the site collection has only recently been added then the label may not yet have been synchronized down.

5.3.2.3. Help

The Help tab displays a list of clue type information, as well as allows you to run the product tour specific to the Taxonomies area.

5.4. Add a Taxonomy

Review the following procedures:
5. Taxonomies

Upload Default Taxonomy

1. In administrative web console, navigate to Taxonomies → Global Settings.
2. Navigate to Loaded Taxonomies, select Add Taxonomies.
3. Select taxonomies that you want to add in the list.

   **NOTE:** Multiple taxonomies selection supported. Clicking the search field enables drop-down list of default taxonomies.
4. Click Load.

5.5. Manage Taxonomies

This section contains information on how to add, merge, back up and delete taxonomies.

Review the following for additional information:

- Create a blank taxonomy
- Importing Taxonomies
- Merge SQL Taxonomies
- Merge SharePoint Taxonomies
- Backup/Delete Taxonomies
- Compare Taxonomies
- Bulk Updates

Create a blank taxonomy

SQL taxonomies reside within the administrative web console database, they are fully functional with the exception of writing metadata back to SharePoint.

To add a SQL taxonomy:

1. Navigate to the Global Settings tab
2. Select the Add button, and finally select the New tile.

Importing Taxonomies

To import an existing taxonomy go to the Global Settings tab, select Add and then choose one of the
import options:

- **SharePoint**—The URL should be set to any site collection within the farm or tenancy, such as: https://netwrix.sharepoint.com. The supplied credentials must have access to both the site collection specified, as well as the termstore (preferably as a term store administrator).

- **Upload**—Imports an XML file directly into the SQL database, large taxonomies will be imported by the background services.

- **Load**—Certain taxonomies are provided out-of-the-box these can be fully used as part of the product or simply used as a reference for regular expression and metadata clues.

**Merge SQL Taxonomies**

SQL taxonomies also be easily merged / updated from the **Global Settings** page. Select the **Update** link for the taxonomy that you wish to update to load the taxonomy merge wizard:

![Merge SQL Taxonomies](image)

Predefined taxonomies can be updated from the latest built-in definition or from an XML file in the standard taxonomy format:

![Update Predefined Taxonomy](image)

The merge operation will automatically add any new terms, update the clues of existing terms, and when enabled delete terms that no longer exist in the new taxonomy definition.
Custom clues can be retained by selecting the option **Retain custom clues**. When enabled any clues not defined as **Predefined** will be retained. The **Predefined** flag can be viewed by selecting the "i" icon for a clue to display the following dialog:

![Details dialog](image)

Any predefined taxonomies that have been previously loaded will show an asterisk indicator when an update is available (post upgrade):

![Asterisk](image)

**NOTE:** The merge operation relies on matching the source definition to the destination definition - utilising the Term Id (GUID). If there are no matching ids then the merge operation will be automatically stopped. In this case the taxonomy should be deleted - and re-imported.

**Merge SharePoint Taxonomies**

SharePoint taxonomies can be merged with the use of the TermStoreManager tool (please see the associated user guide available via documentation downloads).

**Backup/Delete Taxonomies**

Existing taxonomies can be managed via the **Global Settings** tab:
Taxonomies can be exported as XML regardless of the taxonomy type, as well as removed. When removing SharePoint Term Set registrations the source Term Set remains intact - all that is removed is a link to the Term Store.

**Compare Taxonomy definitions**

User can compare current XML taxonomy definition (terms, clues, etc) to an updated/older definition. Comparison matches on the GUID of each term in the source/destination and ignores term movements. The following types of changes are detected: Term additions, term deletions, clue additions, clue deletions, and clue updates.

Do the following:

1. On the **Global Settings** tab, go to **Taxonomies**, select the one you need and click **Compare**.
2. In the **Compare** dialog, select what taxonomy definition to compare to:
   - To compare with the latest predefined definition, click **Yes**.
   - Otherwise, click **No** and browse to the required **comparison file**, i.e. the one that current taxonomy definition will be compared to.
3. Click **Compare** and wait for the process to complete. Examine the results.

**Bulk Updates**

The taxonomy update wizard allows large repetitive changes to be made to taxonomies in bulk. The wizard can be used to:

- **Add Clues**—Create a default standard clue, a default metadata clue, or simply define the clue template to be used.
- **Update Clues**—Update or replace text within the clue text and reference, adjust the score (statically or by percentage), set the local/predefined flags for each clue.
- **Delete Clues**—Remove specific/matching clues.
The wizard is started run by right-clicking a node within the treeview and selecting "Perform Bulk Update". Updates can be performed across the whole taxonomy by right-clicking the root node or scoped to a particular branch by right-clicking the top node of the intended branch:

The wizard will then walk you through performing the update. Each update will allow you to restrict the scope of your change by specifying:

- **Filters**—filters for which terms/clues you wish to update (based on score, clue text, etc).
- **Descendants Limit**—specify how many levels down the update should process within the tree.
- **Exclusions**—specific terms to exclude from the update.

The update can either be performed immediately or in "report-only" mode. When report-only mode is used the scope of changes will be specified to the end-user—the end-user can then choose to commit the update which will perform the changes (or, leave the update if the scope was incorrect).

All updates, report-only or otherwise, can be found under the "Bulk Updates" tab. Updates are queued and processed in the background with the results exposed through this interface.
5.5.1. Managing Term Sets

To manage the term set, select the taxonomy you need, then in the taxonomy tree browse to the required term set and click the Term Management tab on the right.

Then you can work with the tabs you need, including Search, Browse and Working Set tabs.

Review the following for additional information:

- **Documents Movements**
- **Classifications**
- **Calculations**

5.5.2. Multi-User Environments

When several users are maintaining the taxonomy structure simultaneously there is a need to prevent concurrent access to individual classes so that one user’s work is not overwritten by another user working in the same area of the taxonomy.

In order to allow multiple users to work simultaneously we provide a locking facility that allows each user to reserve one or more classes for private editing. When they have finished a batch of work then they can unlock the classes to release.

In order to enable this facility the administrator should “Enable User Locking” under Config → Core → Query Server.

The administrator should also ensure that Anonymous Access is disabled for the administration web application in IIS so that individual Windows identities are available within Taxonomy Manager for locking purposes.

When this facility has been enabled then you will see a Lock Class button in the treeview context menu for all classes:
You can also optionally lock all of its children in a single operation. Once a term is locked the context menu items will change to allow unlocking the selected term, and its children.

Other users will see a closed padlock symbol to indicate the status of the term.

Other users are unable to alter or unlock a term that has been locked by another user. However super-users are also able to Unlock a term.

### 5.6. Search and Filter Taxonomies

When working with taxonomies the hierarchical structure is displayed on the left hand side of the page, allowing for specific terms to be selected and managed. The dropdown list shows all of the taxonomies that are available for management, where appropriate these will be grouped by the SharePoint Term Group.
Right-clicking the tree view nodes provides a number of management options at both the term and termset level including:

- Add Child Term
- Rename Term
- Delete Term
- Re-Classify Term
- Re-Classify Tagged
- Pin Term With Children
- Reuse Terms
- Export CSV

You can also drag-and-drop a node from one location on the tree view to another, once you have dropped the node you can select to either move, copy, or merge the node(s).

Browser rendering restrictions limits the maximum suitable size per level within the tree view at 10,000 terms. Therefore we recommend that the tree view is structured across multiple branches, both for
performance and usability. Once a branch within the taxonomy reaches 10,000 terms the tree view will cap the returned nodes and log a warning to the event logs.

Review the following for additional information:

- Searching for Taxonomy Terms
- "Sync Enabled" Treeview Filter
- Source Filter

**Searching for Taxonomy Terms**

A search facility is provided to locate terms that contain specified text:

Click the magnifying glass icon to the right of the taxonomy dropdown and a new edit box appears where search text may be entered:

"Sync Enabled" Treeview Filter

For SharePoint Term Sets the treeview can optionally be filtered to only show terms that are enabled for synchronisation (configured on the term **Settings** screen).

This setting is session specific and applicable only to the current user:
5. Taxonomies

See Taxonomy Settings for more information.

Source Filter

A filter facility is also provided to restrict all search/browse results to a specific source. Click the source filter link in the top right of the display, then, select a source:

Source Filter: https://conceptssearching.sharepoint.com...

The filter setting can be stored for the session, or just maintained for the browser window.

5.7. Classification Rules (Clues)

Each taxonomy contains a set of terms. Terms are defined by set of configuration rules (also called clues). Clues are used to describe the language found in documents, making these documents belong to a particular topic.

5.7.1. Predefined Classification Rules

The standard taxonomies provided with Netwrix Data Classification include predefined classification rules for personally identifiable information (full name, home address, etc.). They are available in the following languages:

- English
- French
- German
Users can easily extend the out-of-the-box classification rules by adding relevant keywords and terms in other languages.

In addition, there are predefined classification rules for various national identification and registration numbers. These rules typically look for ID patterns supplemented by related keywords for better classification precision.

These rules are provided for the following countries (coverage varies):

- Australia
- Brazil
- Bulgaria
- Canada
- Denmark
- France
- Germany
- Hong Kong
- India
- Italy
5. Taxonomies

- Netherlands
- Singapore
- South Africa
- Spain
- Sweden
- United Kingdom
- USA

5.7.2. Working with Clues

To work with the clues, select the required subnode (terms set) under the taxonomy tree on the left and then select Clues on the right:

- For each clue in the list, you can view and manage its type, score, and other properties.
- To add a new clue, go to the topmost row in the list and specify its properties.

5.7.3. Documents count

Click the Doc Counts link in the top right corner to get the number of documents that match the word / phrase used within the clue:
5.7.4. Suggested Clues

The suggested clues feature facilitates the process of tailoring classification rules in context by offering relevant terms and keywords based on previously indexed file content. This feature is available for all Latin script based languages with increased support for the languages that have support for stemming and/or stop-word analysis:

- Afrikaans
- Danish
- Dutch
- English
- Finnish
- French
- German
- Hungarian
- Italian
- Norwegian
5. Taxonomies

- Spanish
- Portuguese
- Romanian
- Swedish
- Welsh

See also:
- Types of Clues
- Manage Clues

5.7.5. Types of Clues

The following clue types of clues are available, each clue type is described in detail below:

- **Standard Clues**
- **Case-Sensitive Clues**
- **Phrasematch (Wildcard) Clues**
- **Metadata Clues**
- **Phonetic Clues**
- **Regex Clues**
- **Required Terms clue**
- **Term Boost Clues**
- **Language Clues**
- **Static Clues**
- **Hierarchical Clues**

**Standard Clues**

A single word, multi-word concepts or phrases. Use quotes around standard clues to invoke a case-insensitive exact match on entered text, including any punctuation.

**Examples:**

A standard clue matched on a fuzzy basis with word stemming enabled: training will match against: train, training, trains.

A standard clue enclosed in double quotes will be matched on an exact match basis: "Train timetables in the U.K." will match only against: Train timetables in the U.K. (Case-insensitive)

**Case-Sensitive Clues**
A case-sensitive phrase match clue, including any punctuation. There is no need to put double quotes around the text (double quotes at the start and/or end of the text will be removed).

**Phrasematch (Wildcard) Clues**

A phrase match clue that supports the use of '*' and '?' wildcards when matching document text (see [Regex Clues](#) for full REGEX support).

**Metadata Clues**

A clue based on document metadata, with matching based on:

- Exact string matches – Such as: AUTHOR=JOHN SMITH
- Wildcard string matches – Such as: AUTHOR*=john smith*
- Full regex string matches – Such as: AUTHOR^=john.*smith
- Date Range matches – Such as: FIELD > VALUE
- Dynamic Date Range matches – Such as: FIELD > TODAY OR FIELD > TODAY-14 (Matching the last 2 weeks)
- Integer Range matches – Such as FIELD > VALUE or FIELD

Helpers are provided to format metadata clues, to activate the helper simply select the appropriate icon for the desired clue type (numeric, date, and basic): ![Helper Icons]

The date helper supports assisting in the creation of both static and dynamic date clues:

![Create Date Metadata Clue]

Both field and value are case-insensitive for metadata matches.Wildcard matches must include a * character before the equals sign (as shown in the example above).

The following special metadata fields can be used:

**CSE-CONTENTTYPE**

The raw content type, for example:

```
text/*
```
5. Taxonomies

Most applications should use the CSE-TYPE field or the FILE TYPE field (see below) rather than the CSE-CONTENTTYPE field due to the highly variable nature of the raw values.

Examples:
A clue based on PDF documents would look like this
   cse-type = application/pdf
A clue based on a specific author would look like this
   author=john smith

CSE-DOCTYPE
The DocType integer field

CSE-Filename
The document filename (e.g. “Pensions.doc”)

CSE-FILEPATH
The document path not including the filename (e.g. “http://www.bbc.co.uk/sport/”)

CSE-FOLDERS
Used to match folders including sub-folders. For example:
CSE-FOLDERS=http://www.abc.com/jobs/
matches: http://www.abc.com/jobs/123.txt
and also: http://www.abc.com/jobs/UK/123.txt
A clue based on a right truncated path would look like this
CSE-FOLDERS=c:\myfolder\subfolder\
or
CSE-FOLDERS=http://www.abc.com/jobs/
Note that when using cse-Folders with a right-truncated path the path must always end with a slash character.

A clue based on selected folders within the path would look like this
CSE-FOLDERS=myfolder/myfolder2
Note that when using cse-Folders with subfolder matches the value must not begin or end with a slash character.

CSE-FOLDER
Used to match folders without including sub-folders. For example:
CSE-FOLDER=http://www.abc.com/jobs/
matches: http://www.abc.com/jobs/123.txt
does not match: http://www.abc.com/jobs/UK/123.txt

CSE-LASTMODIFIEDDATE
The LastModifiedDate from the collected content in the format “YYYY-MM-DD HH:MM:SS”.
This field can only be matched using the greater than or less than operators, for example:
CSE-LASTMODIFIEDDATE CSE-LASTMODIFIEDDATE > 2010-01-01
Only the date can be specified, not the time.

CSE-LANG
The dominant language of the document, using ISO 639-1 two-letter codes. See Language Detection settings for more information.

CSE-METADATACOLLECTIONONLY
This value will be set to “1” if the document was too large for the NDC index (max 500MB), but was processed using metadata only.

CSE-PAGETITLE
The Title extracted from the document itself.

CSE-TEXTLENGTH
The length of the plain text extracted from the document, in characters.
This field can only be matched using the equals, greater than or less than operators, for example:
CSE-TEXTLENGTH = 50000
CSE-TEXTLENGTH > 50000
CSE-TEXTLENGTH

CSE-TITLE
The Title extracted from metadata.

CSE-URL
The document URL, including the filename (e.g. “http://www.bbc.co.uk/sport/Pensions.doc”)

FILE TYPE
The short normalised content type, always one of the following:
Adobe PDF files:
PDF
Corel WordPerfect files:
WPD
Microsoft Excel files:
XLS
XLSX
Microsoft Outlook MSG files:
MSG
Microsoft PowerPoint files:
PPT
PPTX
Microsoft Rich Text Format files:
RTF
Microsoft Word files:
DOC
DOCX
Text files (including HTML, XML, CSV, etc.):
TXT
HTML
XML
All other file types
OTHER

FILE SIZE
The length of the document, in bytes.
This field can be matched using the equal, greater than or less than operators, for example:
FILE SIZE = 10000
FILE SIZE FILE SIZE > 10000
The Modified date from the document metadata in the format “YYYY-MM-DD HH:MM:SS”.
This field can be matched using the equal, greater than or less than operators, for example:
MODIFIED = 2010-01-01
MODIFIED MODIFIED > 2010-01-01
Only the date can be specified, not the time.

Phonetic Clues
A case-insensitive fuzzy/phonetic phrase match clue. Phonetic clues ignore all non alphanumeric
characters. Words that contain no digits are matched using a phonetic algorithm so that words that sound similar will be matched. Phonetic clues do not use word stemming in the matching process.

For example, the following clue:

Intelligence Organisations in the Middle East

Would match any of the following:

- Intelligence Organizations in the Middle East
- Intelligence Organisations in the Middle-East
- Intelligence organisations, in the “middle east”.

But not any of the following:

- Intelligence Organisations located in the Middle East
- Intelligence Organisations in the Mid-East

Regex Clues

A Regular Expression matching clue – by default this is run across all document text and metadata. Regular expression clues are run in a case-sensitive manner by default. You can optionally enable the "Case-Insensitive Regex Processing" mode, this setting can be found in Config -> Classifier.

Definitions of the required syntax for regular expressions can be found in many places, including Microsoft: Regular Expression Syntax.

The following example clue matches US Social Security Numbers found anywhere in the document text:

```
[/,,/.,/=,\s]((?!000)\[0-6\]d\{2\}|7[0-6]d\|77[0-2])-(\(?!00\)d\{2\})-(\(?!0000\)d\{4\})[/,,/.,\s]
```

This sample clue ensures that:

- The SSN must consist of 11 characters in this format: NNN-NN-NNNN
- The SSN must be preceded by a white space or a dot or a comma or an equals sign.
- The SSN must be followed by a white space or a dot or a comma.
- The two hyphens must be present.
- None of the three sections can be equal to zero.
- The first section must be in the range 001 – 772

Any regular expression matches found will be extracted and added to the NDC index automatically. For example, if we have a document that contains this text:

*Here is one SSN: 407-54-8831*

*And here is another 407-54-8832 in the middle of this sentence.*

Then the following metadata entries will be generated automatically:
5. Taxonomies

These can easily be viewed within the document “Info” popup on the “Metadata” tab (filtered to Regex values). The automatically generated metadata field name is a combination of the term name prefixed with “Regex-“.

Regular Expression Result Validation

In some cases it may be necessary to assign certain requirements on the result of the regular expression. This is particularly relevant for expressions that may include false positives such as social security numbers (simple pattern) or credit card numbers (sample data). The classification engine includes a number of post match validation steps:

- **Exclusion Patterns**—Provides the ability to exclude a match based upon an exclusion pattern (exclude sample data etc). Exclusion patterns can be added by selecting the “Exclusions” link. If any exclusion rule is matched the regular expression result will be discarded.

  **TIP:** Hit count based regular expression clue exclusions — restrict whether a regular expression clue should match based upon the number of unique matches found against the regular expression. I.E, a regex to match any number against the text: “1 1 1 2 3 4” - has 4 hits, 4 unique numbers.

- **Validation Checks**—Certain patterns correspond to particular validation checks (such as credit card numbers, international bank account numbers etc). Currently supported checks include:
  - Mod 97/10
  - Luhn
  - Verhoeff

  *To add a validation check:*
  1. Select the **Exclusions** link for the desired clue
  2. Click **Add**
  3. Select the desired check from the drop down selection
  4. Click **Save**.

  *If any validation check fails then the regular expression result will be discarded.*

- **Proximity Matches**—Provides the ability to include/exclude regular expression matches based upon the existence of text before or after the regular expression match. Matches can be added by selecting the “Proximity Matches” link. Matches are processed as follows:
  - If any ‘Exclude’ match passes then the regular expression result will be discarded
  - If no ‘Include’ matches exist – or, at least one ‘Include’ match passes then the regular expression result will be considered valid
NOTE: This functionality is only available when utilising classification Engine v2. The additional settings are also not currently available in SharePoint Term Sets (but can be linked via Term Boosts).

**Required Terms clue**

The **Required Term** clue type can be used to require another class to be classified as a pre-requisite for this class. This is most often used when the children of a class require the parent to also be classified.

The valid entries for this type of clue are:

- Parent
- Grandparent
- Any specific term in any taxonomy

A tree view control makes selecting the required class easy:

For example, suppose that we have a topic **Pensions** with two children:

- Pensions
  - USA
  - Canada

The purpose of the two child classes is to identify documents that are about pensions in the USA or about pensions in Canada. Rather than add clues to identify pensions documents to the children you can simply require documents to be about **Pensions** by using a Required Class clue type.

**Term Boost Clues**

The **Term Boost** clue type can be used to specify that a Class Score is to be boosted from another term. This is most often used when a complex class is implemented using several child (or even grandchild) classes.

A tree view control makes selection of boosting classes easy.
The score may be entered as a number (if a fixed boost is required regardless of the source term's score) or as a percentage (if the boost score is to be calculated as a percentage of the source term's score).

When referencing a specific node it is also possible to include one or more levels of that node's descendants. At classification time if the referenced node or any of its descendants (up to the configured level) reach their threshold then the term boost will be applied.

**Language Clues**

The language clue type can be used to require documents to be written primarily in a specified language as a filter on classification.

For example, if you create a new class and want documents to be classified only if they are written in a Scandinavian language then you would create a Language clue, like this:

**Static Clues**

The static clue applies a score to the class without any pre-conditions, this can be useful when creating NOT functionality.

For example:

If you want to classify any document where a word does NOT exist (such as *Pensions*), you could first add a static clue with a score of 50, and then add a standard clue looking for *Pensions* with a negative score (-50).

**Hierarchical Clues**

Hierarchical clues support a parent-child clue hierarchy, if the child clues achieve the parent clue threshold then the hierarchical score will be applied.

This can be useful when you only want to apply a score if two or more conditions to match, or perhaps to only apply a small static score if a word appears X times within a document.
5.7.6. Adding a Clue

To add a new clue, go to the topmost row in the list and specify clue properties, as explained below:

- Type
- Clue (rule body)
- Score
- Is Mandatory

When ready, click **Insert** on the right.

5.7.6.1. Clue Body

When specifying the clue body, consider exact matching and stemming explained below.
5.7.6.1.1. Exact Matching

There may be any number of words up to a maximum of 200 characters per clue. However, most clues will consist of one, two or three words.

Use double quotes around phrases to invoke exact phrase matching.

5.7.6.1.2. Stemming

Word stemming simplifies classification rules by automatically matching inflected word forms using a single keyword clue. This can be useful to identify how a clue will be implemented by the classification engine. Stemming is supported for the following languages:

- Dutch
- English
- French
- German
- Hungarian
- Spanish
- Portuguese

Hovering over a standard clue will show the stemmed version of the word / compound term.

**Example:** A class called *Global Warming* may have the following clues:

- Global Warming
- Greenhouse Gases
- CO2 Emissions
- Pollution

To disable stemming, use double quotes around single words.

5.7.6.2. Score

Scores are expressed as percentages of the threshold. For example, if the threshold is 50 then:

- 50 = guarantees that this term alone will be sufficient to classify the document
- 25 = this term will get half way to the target
- 10 = this term is of low importance but its presence should boost a document score
5. Taxonomies

- 0 = zero weight – use to disable a clue
- -10 = this term is a small negative indicator
- -50 = this term is a strong negative indicator
- -1000 = the presence of this term should force the document to not be classified

Higher scores indicate a stronger association with the topic.

- **Example 1**: *Global Warming* with a score of 50 will cause a document with this concept to be matched.
- **Example 2**: *Pollution* with a score of 20 (on its own) will not be sufficient to cause the document as being about global warming.

Consider that clues can also be assigned a negative value, which will prevent incorrect associations.

- **Example 3**: *Noise pollution* should not be associated with *Global Warming*. So *Noise pollution* would be added with a negative value.

### 5.7.6.3. Mandatory Clues

You can use the **Mandatory** checkbox to indicate that a clue is required, i.e. a document cannot be classified against a category unless it matches all of the mandatory clues.

The mandatory clue selector is denoted by the * icon:

![Mandatory Icon](image)

### 5.7.6.4. Using the Local Option

In some cases, a further option will be available per clue: “Is Local?”. This option allows the user to restrict a clue purely to the current Term Set.

**NOTE**: This option is only available for reused terms (SharePoint Term Sets).

![Local? Icon](image)

- Once this option is selected, it will not be possible to amend the clue from any other Term Set that contains the re-used Term.
- If you want to share the Term across all Term Sets again, clear the option from the Term Set in which it was originally enabled.
5.7.6.5. Using Synonyms (SQL taxonomies only)

**NOTE:** The Synonyms link is only available for the clues in SQL taxonomies.

The Synonyms link can be used to enter synonym definitions.

In general, the use of this facility is not recommended. The preferred approach is to enter each synonym as separate clues. Entering each synonym as separate clues will generally result in more accurate scoring and therefore to better classification results.

5.7.7. Manage Clues

This section describes how you can edit, import, move and delete clues for the selected taxonomy term set.

- To delete a clue, select the checkbox next to it and click **Delete**.
- To edit a clue, select it from the list and click **Edit** link on the right. Then you can modify clue type and provide the appropriate settings. See **Types of Clues** for details. To see how the edits will take effect, click **Preview** on the right. To apply edits, click **Update**.
- To modify all selected clues, see **Bulk Edit**
- For bulk import of clues from an Excel Spreadsheet, click **Bulk Insert**. See **Bulk Import**.
- To move or copy the clue to another term, select it from the list and click **Copy/Move**. Then select the destination term and click the button you need (**Move** or **Copy**).

See also:

- **Types of Clues**
- **Adding a Clue**

5.7.7.1. Bulk Edit

The **Bulk Edit** link can be used to make changes to several clues at one time:
When this link is used the form changes into a grid editor and many values can be changed and saved in a single operation. To alter the Mandatory or Is Local settings for all terms quickly simply click the header text to toggle all checkboxes between enabled / disabled.

It is also possible to preview the changes made whilst in the bulk editor. The Preview functionality provides an indication of the number of documents affected, and the resultant score change:
5.7.7.2. Bulk Import

Clues can also be imported in bulk from an Excel Spreadsheet (or input in bulk manually). The spreadsheet should contain 3 columns: Type (Standard, Case-Sensitive, Wildcard Phrasematch or Metadata), Clue Text and Score:

```
<table>
<thead>
<tr>
<th>Type</th>
<th>Clue</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Standard</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Standard</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Standard</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Standard</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Standard</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Standard</td>
<td></td>
<td>50</td>
</tr>
</tbody>
</table>
```

Simply paste the spreadsheet content here.

The **Bulk Insert** link is available on the **Clues** tab below the main entry grid.

5.7.8. Search Documents by Clue

You can search for documents based on the class clues. For that, click on the name of any single clue in the clue list in the management console (or even any suggested clue), go to the **Search** tab and configure search settings.
1. Set up the following properties that will be considered a basis for the search:
   
   - Clue type - select the required value from the **Type** list.
   
   - Clue itself (clue body) - enter the required keyword or phrase in the **Find** field.

   **NOTE:** See [Classification Rules (Clues)](classification-rules-clues) for more information.

2. To restrict the search further, you can either add a **URL** filter, or add a custom filter by clicking **Add custom filter** link. This can be helpful when evaluating the usefulness of a clue by quickly examining its usage within the corpus. Consider the following:

   - The URL filter must end on a folder boundary.

   - Use custom filter to specify a number of complex filters: boolean, datetime and numeric.

   **NOTE:** Full description of all filters can be found in the API Reference Guide.

3. To view how recent changes to the term will affect the document classifications, select **Show document movements**. As a result, the “movement” of the document since the last classification will be shown. Possible scenarios are:

   - **↑** Document remains classified with a higher score
   
   - **↓** Document remains classified, but with a lower score

   - **=** Document remains classified. Score does not change

   - **✓** Document will become classified

   - **✗** Document either stays or becomes un-classified

OR
5.7.9. Browse

To view the documents classified for each term, click on the **Browse** tab. This will display a list of documents achieving the minimum score set for classification in the term. See [Classification Rules (Clues)](none) for more information.

**NOTE:** This list will include the current classification status of each document and any changes made to the class, since the last classification, are not taken into account.

The document text will be highlighted based upon the clues configured for the term. Highlighting will include regular expression matches when configured (**Config**→**Query Server**→**Enable Regex Browse Highlighting (Advanced)).

**NOTE:** If a new class is selected in the treeview menu, the view will remain in "Browse" mode and will show the documents for the selected class.

You can use the **Browse** function to:

- Identify documents that are receiving a score, but are "missing" being classified because they do not quite reach the terms threshold. For example, changing the mode to "Near Misses <20%" for a term with a threshold of 50, will find any documents that scored 40 or more, but did not reach the threshold.
 Identify low scoring documents that are only just reaching the classification threshold. For example, changing the mode to "Low Scoring Documents <20%" for a term with a threshold of 50 will find any documents that scored between 50 and 60.

To restrict the browsing scope, you can either add a URL filter, or add a custom filter, as well as select to show document movements. These options are configured in the same way as for Search Documents by Clue.

5.7.10. Export Search Results

Search / Browse results can be exported quickly and easily by selecting the either of the export options below the search results:

- Quick Export
- Queued Reports

If there are less than 1000 results, or you wish to have access to the results immediately, you can select the Quick Export option (light icon).

Alternatively the export results will be created in the background, and made available later view the Queued Reports area. A notification can be sent to an email group upon the completion of report processing, when selected:
5.8. Suggestions

Clues can be used to statistically produce a list of suggested clues that can be assigned to the term.

<table>
<thead>
<tr>
<th>Type</th>
<th>Clue</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>Natural resources</td>
<td>40</td>
</tr>
<tr>
<td>Standard</td>
<td>environmental protection</td>
<td>30</td>
</tr>
<tr>
<td>Standard</td>
<td>climate</td>
<td>20</td>
</tr>
<tr>
<td>Standard</td>
<td>pollution</td>
<td>20</td>
</tr>
<tr>
<td>Standard</td>
<td>conservation</td>
<td>10</td>
</tr>
<tr>
<td>Standard</td>
<td>protection</td>
<td>10</td>
</tr>
<tr>
<td>Standard</td>
<td>Wildlife</td>
<td>10</td>
</tr>
<tr>
<td>Standard</td>
<td>Environment</td>
<td>10</td>
</tr>
</tbody>
</table>

Clues can be suggested for a term via the following methods:

- **Suggest Clues for whole term:** Click on the **Suggest Clues for class** link under the class heading to produce a list of suggestions, based on all existing clues in the class.
- **Single Clue:** Click on the **Suggest** link against each clue to produce a list of suggestions, based on only this clue.
- **Class Document:** Click on the **Suggest** link against each class document to produce a list of suggestions, based on the document.

Once the list of suggested clues has been generated they can be selected and added to the term clues:

**NOTE:** Changes made to a class will have no effect unless documents are re-classified.
The clue type can be set to one of the following:

- Standard
- Case-Sensitive
- Phonetic
- Create Tree Node

**NOTE:** If Create Tree Node is selected then these topics shall be added as children of the currently selected node in the taxonomy structure.

### 5.9. Working Set

A Working Set of documents can be defined and used to test the accuracy of classification rules against a controlled set of documents. The Working Set is mode can be selected in the [Core Configuration](#).

If Class Level is selected then a different Working Set can be defined for every class. If Taxonomy Level is selected then the same Working Set will be used for all classes.

Documents can be added to the Working Set from the Search or Browse tabs by using the Add to Working Set links:

The following facilities are available:

- [Documents Movements](#)
- [Classifications](#)
- [Calculations](#)
5.10. Related

The Related tab allows you to view and modify the non-hierarchical relationships between preferred terms. This tab will only appear if the taxonomy is in SQL, as the SharePoint Term Store does not support this functionality.

When a term is located in multiple branches of the taxonomy (a polyhierarchical taxonomy) – the Related tab will also display each of the locations to allow you to jump to the specific branch.

5.11. Additional Configuration

This section contains information on additional and / or optional tabs. Review the following for additional information:

<table>
<thead>
<tr>
<th>Tab</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph</td>
<td>The Graph tab shows a graphical representation of classification intersection points.</td>
</tr>
</tbody>
</table>
In the example above 6721 documents are tagged with "Medium (100kb-1Mb)", 1254 of these documents are also tagged with "HTML". It's also possible to see that there are 3517 documents that are tagged with both "HTML" and "English" (highlighted by the dashed links).

**Info**

The **Info** tab displays the term description (aka Scope Notes) for each preferred term. The **Description** field is often populated automatically when an external taxonomy is imported automatically using the Scope Notes.

**Logs**

All changes made to a term are recorded. The change history may be viewed from the Logs Tab:

When auto-classifications are amended in SharePoint the user edits are recorded in the SQL database, these can later be reviewed to identify terms that require review:
5. Taxonomies

### User Edits

<table>
<thead>
<tr>
<th>Tab</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User Edits</strong></td>
<td>An optional interface can be enabled to allow users to suggest new terms for the termset hierarchy (<a href="http://netwrixdataclassificationserver/conceptQS/Taxonomies/TermSuggest.aspx">http://netwrixdataclassificationserver/conceptQS/Taxonomies/TermSuggest.aspx</a>). Suggestions can trigger automatic notifications to taxonomy administrators, as well as being recorded in the database for later review on the &quot;User Suggestions&quot; tab:</td>
</tr>
</tbody>
</table>

#### User Suggestions

<table>
<thead>
<tr>
<th>Term</th>
<th>Suggested Path</th>
<th>Requester</th>
<th>Email Address</th>
<th>Reason</th>
<th>Request Date</th>
<th>Processed Date</th>
<th>Processed By</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>DICOM</td>
<td>Root</td>
<td>John Smith</td>
<td><a href="mailto:john@demo.com">john@demo.com</a></td>
<td>Identification of DICOM files</td>
<td>2019-03-03 13:04:20</td>
<td>-</td>
<td>-</td>
<td>Accept</td>
</tr>
</tbody>
</table>
6. Workflows

6.1. Understanding Workflows

A workflow allows you to configure an automated action that will be performed on a document, following a classification decision. For example:

- Send an email message to personnel in charge
- Move or copy a document from one location to another, and many others.

To set up a workflow, you need to do the following:

- Specify conditions, defining the classification decisions that this workflow will act upon.
- Configure rules that will trigger your workflow actions.
- Select actions that will take place when one or more rule conditions are met.

Looking for real-life use cases and walk through examples? Check out Netwrix training materials. Go the Netwrix website to find out how you can easily reduce the exposure of your sensitive data.

See next:

- Managing Workflows
- Workflow Actions

6.2. Managing Workflows

Authorized users can create, modify or delete automated workflows that apply to the certain content. For that, in the administrative web console select Workflows from the top menu and go to the Workflows tab.

![Workflows Table]

NOTE: To manage the automated workflows, users require sufficient access rights that are assigned based on either their Windows identity or using non-Windows based access controls. See "Users" for details on rights and permissions.

- Click Copy if you want to copy the list content to the clipboard.
- You can also export the list to CSV or XLSX file.
By default, the number of list items displayed (Page Size) is set to 10. Modify this setting as necessary.

To delete all workflows from a certain scope, select the corresponding list item and click **Delete**.

**To create a Workflow**

To create an automated workflow for certain type of documents, you can use the **Add Workflow** wizard or **Advanced** dialogs.

See next:

- [Create a Workflow using Add Workflow Wizard](#)
- [Configure a Workflow using Advanced dialog](#)

**To modify or delete a Workflow**

To modify a workflow, follow the steps described in the **Edit Workflow settings** section.

To delete a workflow, follow the steps described in the **Delete Workflow** section.

**To clone, enable or rename a Workflow**

1. Click the link in the **Name** column for the required workflow (e.g. **Global** for **Google Drive** in the figure below):

2. This will open the list of workflows for selected scope. You can sort the list by **Details** (workflow action) or by **Active** (workflow state) field.

3. Select one or several workflows you need.

4. To **Disable** or **Enable** the workflow, use the corresponding button above or link on the right. Workflow state (**Active** field) will change accordingly.

5. If you want to create a copy of selected workflow, with all associated actions and conditions, click **Clone**, then enter the scope (group) and name for the new workflow.
NOTE: Workflows within a generic group (scope) are cloned within the same group, source-specific workflows can be copied within any groups of the same type. The clone workflow will be disabled by default.

To provide another name to a workflow, select it from the list and click Rename.

NOTE: Workflow names must be unique within the group (scope).
6.2.1. Create a Workflow using Add Workflow Wizard

To create an automated workflow for certain type of documents, you can use the Add Workflow wizard or Advanced dialogs.

To launch the Add Workflow wizard:

1. In the administrative web console, select Workflows from the top menu.
2. Click the Workflows tab.
3. Click the Add button in the upper right corner.
4. In the dialog displayed, click the Launch Wizard button.

See next:

- **Step 1. Select Content Type**
- **Step 2. Select Action**
- **Step 3. Specify Conditions for Processing**
- **Step 4. Enter Name and Review Settings**

**NOTE:** Once created you will be able to modify the workflow using the Advanced dialog.

Alternatively, take steps 1-3 from the procedure above, then in the Add Workflow dialog click Advanced. See **Configure a Workflow using Advanced dialog**.
6.2.1.1. Step 1. Select Content Type

At the first step of the wizard, select the type of content your workflow will process, and specify which content sources of that type should be included in processing.

1. From the drop-down list, select what type of documents you want this workflow to target:
   - To apply the workflow to all types of content, selecting All types.
   - Otherwise, select what type of content you want to be included in the workflow:
     - Exchange
     - File
     - Google Drive
     - SQL

2. Then specify which source of content you want to process. You can select All sources, or select the one you need.

   ![Add Workflow](Image)

   Click Next to proceed.

   See also: Content Sources.
6.2.1.2. Step 2. Select Action

After you select the required type of content source, you will be offered the number of automated actions available for such content, for example, send an alert by email or update document metadata, etc.

Click the action you need and configure the necessary settings. For details, see *Available Actions*.

When finished, proceed to the next step.
6.2.1.3. Step 3. Specify Conditions for Processing

At this step, you can specify whether workflow actions should be performed with the classified documents only, or with any documents from the content source, etc.

The following options are available:

- **Any Document** — with this option selected, the workflow will be applied to all documents in the specified content source
- **Any Classified Document** — with this option selected, the workflow will be applied to the documents in the specified source if they were tagged by any classification
- **Specific Classification** — with this option selected, you need to specify whether to apply the workflow to the classified or non-classified documents
  - To process only documents classified by specific classification, select **Classified** (this will act as including filter)
  - To process only non-classified documents, select **Not Classified**.

If you have selected any of the **Specific Classification** variants, you should then specify taxonomy terms that will be applied to filter out the documents for your workflow.

**To configure terms**

1. In the **Select Term** field, click the tag icon.
2. In the **Details** dialog, specify filter settings to use when filtering out the documents:
   a. **Taxonomy** - select what classification taxonomy from the existing ones should be used.
   b. **All Terms** - select this option if you want to filter by all terms in the taxonomy. If this option is cleared, then after selecting the necessary taxonomy, you will be presented the list of its terms. Select the one you plan to use for filtering.
NOTE: Multiple selection is not supported: to configure several filter values, you should repeat this procedure for each filter value you need.

C. **Include Children** - select this option if needed.

3. Finally, click **OK** to save the settings and close the dialog.

Then verify that configured filters are displayed properly:

- Including filters (i.e. instructing to include documents with classification tag you selected) are colored blue:

![Add Workflow](image1)

- Excluding filters (i.e. instructing to include documents without classification tag you selected) are colored red:

![Add Workflow](image2)

If you have selected more than one filter, you will be prompted what logic should be used when applying the filters:

- To apply **AND** logic (i.e. the document must meet all filtering conditions), select **All**.
- To apply **OR** logic (i.e. the document must meet any of the filtering conditions), select **Any**.
6.2.1.3.1. Example 1. Include All Files Classified as PDF

For example, you want your workflow to process all PDF files from the selected content source. Do the following:

1. Select Specific Classification option.
2. Select Classified.
3. Click the tags icon in the Select Term field on the right.
4. In the Details dialog, from the Taxonomy list select File Type.
5. Then from the list of file types select PDF and click OK.

After you get back to the wizard, the PDF filter will appear under the Classified option, colored blue (indicating this filter is including).

6.2.1.3.2. Example 2. Exclude HTML and XML Files

For example, you want your workflow to process all classified documents from the selected content source, except HTML and XML files. Do the following:

1. Select Specific Classification option.
2. Select Not Classified.
3. Click the tags icon in the Select Term field.
4. In the Details dialog, from the Taxonomy list select File Type.
5. Then from the list of file types select HTML and click OK.
6. After you get back to the wizard, check that the PDF filter is shown colored red (indicating this filter is excluding).
7. Repeat steps 3-5 for the XML file type.

8. Under the **Require all conditions?** prompt select **Any** — for OR logic to be applied, so that any HTML or XML file should be excluded (in other words, the workflow will be applied only to the files not classified as HTML or XML).

9. Finally, click **Next** to proceed.
6. Workflows

6.2.1.4. Step 4. Enter Name and Review Settings

At this step, you need to provide workflow name, review its settings, and disable or enable the workflow (to start immediate processing). Do the following:

1. Enter workflow name. It should contain at least 3 characters.
2. Review the workflow settings you have configured at the previous steps.
3. If you want the documents to be processed immediately after you finish the wizard, select **Enabled** option. Otherwise, you can select **Disabled** and change this settings later on using the UI.

**NOTE:** Documents that have already been classified will be re-classified before applying this automated workflow.

![Add Workflow dialog](image)

When finished, click **Add** to close the wizard. Your new workflow will be added to the list on the **Workflows** tab:
6.2.2. Configure a Workflow using Advanced dialog

This section contains information on how to add or edit workflows using the Advanced dialog window.

To configure a workflow:

1. On the Workflow tab, click Add and in the dialog displayed click Advanced.
2. Specify Name for the workflow.
3. From the Type drop-down list, select the type of content your workflow will apply to.
4. Click Add.

1. Then you need to configure document processing rules. For each rule, you should set up rule conditions and rule actions. Also, specify how the workflow should be processed with regards to rules.
   - Specifying Rule Conditions
   - Specifying Rule Actions
Other Rule Settings

To apply pre-conditions (they will be used before rule processing starts), see Specifying Workflow Conditions

6.2.2.1. Specifying Rule Conditions

1. In the corresponding section on the Rule tab, click Edit on the right. The Edit Rule Conditions dialog will be displayed.

2. From the Mode list, select how the conditions should be applied.

![Edit Rule Conditions dialog](image)

The following options are available:

- **Any Document** — with this option selected, the workflow will be applied to all documents in the specified content source

- **Any Classified Document** — with this option selected, the workflow will be applied to the documents in the specified source if they were tagged by any classification

- **Specific Classification** — with this option selected, you need to specify whether to apply the workflow to the classified or non-classified documents
  - To process only documents classified by specific classification, select Classified (this will act as including filter)
  - To process only non-classified documents, select Not Classified.

If you have selected any of the Specific Classification variants, you should then specify taxonomy terms that will be applied to filter out the documents for your workflow.

To configure terms

1. In the Select Term field, click the tag icon.
2. In the **Details** dialog, specify filter settings to use when filtering out the documents:
   
   **a. Taxonomy** - select what classification taxonomy from the existing ones should be used.

   **b. All Terms** - select this option if you want to filter by all terms in the taxonomy. If this option is cleared, then after selecting the necessary taxonomy, you will be presented the list of its terms. Select the one you plan to use for filtering.

   **NOTE:** Multiple selection is not supported: to configure several filter values, you should repeat this procedure for each filter value you need.

   **c. Include Children** - select this option if needed.

3. Finally, click **OK** to save the settings and close the dialog.

3. You can specify what logic should be used when applying the filtering terms:
   
   - To apply AND logic (i.e. the document must match all filters), select **Require all conditions be met**.
   
   - Otherwise, OR logic will be used (i.e. the document must meet any of the filtering conditions).

4. Make sure the filtering term is displayed in the **Edit Rule Conditions** window with blue color. Click **Save**.

The configured rule condition will appear in the **Rule Conditions** section on the **Rule** tab.

---

**Example**

If you want to apply the rule to all documents classified as Visa cardholder data using PCI DSS taxonomy, configure the rule condition as follows:

1. From the **Mode** list select **Specific Conditions**.

2. Select **Classified** option.

3. In the **Select Term**, click the tag icon.
4. In the **Details** window, from the **Taxonomy** list select **PCI DSS**.

5. In the tags hierarchy, select **Visa** and click **OK**.

![Details window with PCI DSS selected]

Make sure the filtering term is displayed in the Edit Rule Conditions window with blue color. Click **Save**. The configured rule condition will appear in the **Rule Conditions** section on the Rule tab.

### 6.2.2.2. Specifying Rule Actions

1. In the corresponding section on the Rule tab, click **Add** on the right. The **Add Action** dialog will be displayed.

2. From the **Action Type** list, select the action you want to apply to the documents that match rule conditions. For details, see **Workflow Actions**.

3. Click **Save**.

![Add Action dialog with email alert configuration]
6.2.2.3. Other Rule Settings

On the Rule tab, you can also manage the rule, as follows:

- Add another rule, clicking the '+' sign.
- Enable or disable this rule, selecting or clearing the Enabled check box in the top right corner.
- Specify how rule application will affect workflow processing. Possible options are:
  - Processing stops if this rule is run
  - Processing stops if any action fails
- Edit rule conditions.
- Copy or delete the current rule.
- Copy rule configuration as text, CSV or XLSX file.
- Add, Edit or Delete rule actions.

If multiple rule actions have been configured, they will be processed in the order listed. Use the red down arrow or green up-arrow to change the processing sequence as required:

6.2.2.4. Specifying Workflow Conditions

You can narrow the initial workflow scope. For that, specify the conditions that document should match to be processed by the workflow.

1. Go to the Conditions tab. By default, the Workflow Conditions list includes none, that is, current workflow will consider any document; actual filtering conditions will be applied by the rule (rule conditions).
1. Click **Edit** to open **Edit Workflow Conditions** dialog.

2. Select the option you need from the **Mode** list. The next steps are similar to those described in **Specifying Rule Conditions**.
6.2.3. Edit Workflow settings

To edit the workflow settings, do the following:

1. On the Workflows tab, click the row that contains the required workflow.
2. In the list of workflows displayed, click the one you need.
3. You will be forwarded to the configuration window where you can modify workflow conditions, rule conditions and actions, as described in the Configure a Workflow using Advanced dialog section.
6.2.4. Delete Workflow

You can delete a single workflow or a group of workflows within the scope (Global or other):

- To delete all workflows in the scope, in the Workflows list select the necessary Name (scope) and click Delete on the right.

- To delete specific workflow, do the following:
  1. In the list of workflows, locate the workflow you need.

    **TIP:** You can use Search in the upper right corner of the window.
  2. Click the link in the Name column for the required workflow (Global for Google Drive in the figure below):

3. This will open the list of workflows for selected scope. Select the workflow you need and click Delete.
6.3. Workflow Actions

Actions are automated operation to be performed with the documents when rule conditions are triggered. There are two types of workflow actions:

- **Generic actions** available for any type of document. These are:
  - Email Alert
  - Migration
  - Apply Additional Classification

- **Source-specific actions** described in the corresponding sections of this guide. See [Available Actions](#).

Workflow actions are executed at the final stage of the document processing.

See next:

6.3.1. Available Actions

This section lists workflow actions available for the certain content source types.

<table>
<thead>
<tr>
<th>Content source type</th>
<th>Available actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchange</td>
<td>Email Alert</td>
</tr>
<tr>
<td></td>
<td>Migrate Document</td>
</tr>
<tr>
<td></td>
<td>Apply Additional Classification</td>
</tr>
<tr>
<td></td>
<td>Advanced Actions for Exchange*: delete email, move email</td>
</tr>
<tr>
<td>File System</td>
<td>Email Alert</td>
</tr>
<tr>
<td></td>
<td>Migrate Document</td>
</tr>
<tr>
<td></td>
<td>Apply Additional Classification</td>
</tr>
<tr>
<td></td>
<td>Advanced Actions for File System*: update permissions, add/remove MIP label</td>
</tr>
<tr>
<td>Google Drive</td>
<td>Email Alert</td>
</tr>
<tr>
<td></td>
<td>Migrate Document</td>
</tr>
<tr>
<td></td>
<td>Apply Additional Classification</td>
</tr>
<tr>
<td>SharePoint</td>
<td>Email Alert</td>
</tr>
<tr>
<td></td>
<td>Migrate Document</td>
</tr>
<tr>
<td></td>
<td>Apply Additional Classification</td>
</tr>
</tbody>
</table>
6. Workflows

<table>
<thead>
<tr>
<th>Content source type</th>
<th>Available actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced Actions for SharePoint *: send classification value, filtered targeted meta update, write/remove O365 label, copy/move document</td>
<td></td>
</tr>
<tr>
<td>SQL and other databases</td>
<td>Email Alert</td>
</tr>
<tr>
<td></td>
<td>Migrate Document</td>
</tr>
<tr>
<td></td>
<td>Apply Additional Classification</td>
</tr>
</tbody>
</table>

* — these actions can be only configured using the Advanced UI dialog window.

6.3.1.1. Email Alert

This action sends an email to the list of provided email address(es). When running the Workflow wizard and having selected Email Alert as an action, you will be prompted to configure the related settings.

In the case where the Workflow is configured against a SharePoint source / group (or, the generic “All Sources” for SharePoint) the action will optionally support a dynamic recipient selection against either the creator or last modifier of the document (provided by the SharePoint document metadata).

Specify the following:

<table>
<thead>
<tr>
<th>Field</th>
<th>Settings to specify</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific recipients</td>
<td>Specify email address to send the alert to. To enter multiple recipient, click + on the right.</td>
</tr>
<tr>
<td>Who should the email be sent from?</td>
<td>Specify email sender and SMTP server settings. You can select a pre-configured SMTP server (if any), or specify new connection parameters by clicking the + on the right — then in the Email Server Details dialog enter the following:</td>
</tr>
</tbody>
</table>
### Field Settings to specify

- **Host**—Enter your SMTP server address. It can be your company's Exchange server or any public mail server (e.g., Gmail, Yahoo).

- **Port**—Specify your SMTP server port number.

- **Use SSL**—Select this checkbox if your SMTP server requires SSL to be enabled.

- **From Email**—Enter the address that will appear in the From field.

- **Username**—Enter a user name for the SMTP authentication.

- **Password**—Enter a password for the SMTP authentication.

**NOTE:** It is recommended to use Test Configuration Settings option. The system will send a test message to the specified email address and inform you if any problems are detected.

---

![Email Server Details](image-url)

When finished, click **Save** to close the dialog and return to email action settings.
6. Workflows

<table>
<thead>
<tr>
<th>Field</th>
<th>Settings to specify</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email Subject</td>
<td>Specify the template for email subject. The template can contain dynamic values that will be obtained from the crawled content (e.g. <code>[cs:PageUrl]</code>). <strong>TIP:</strong> You can select the corresponding fields from Add a Merge Field list on the right.</td>
</tr>
<tr>
<td>Email Body Template</td>
<td>Specify the template for email body. The template can contain dynamic values that will be obtained from the crawled content (e.g. <code>[cs:PageUrl]</code>). <strong>TIP:</strong> You can select the corresponding fields from Add a Merge Field list on the right.</td>
</tr>
</tbody>
</table>

To modify action settings for the certain workflow, select the workflow and use the Advanced UI window, as described in the Modify Email Alert action settings section.

### 6.3.1.1.1. Modify Email Alert action settings

To modify Email Alert action settings using the Advanced interface:

1. In administrative web console, navigate to Workflows and select the workflow you want to configure email alert for.
2. Click the workflow, then click Add next to Rule Actions.
3. In the Add Action dialog, select Email Alert section in the Action Type list.

Specify the following settings:

<table>
<thead>
<tr>
<th>Field</th>
<th>Setting to specify</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email Address</td>
<td>Specify email recipients. You can enter multiple static email addresses. <strong>NOTE:</strong> Dynamic configurations will use the 'Document Modified/Created By' metadata value, looking up the user's email address from Active Directory where appropriate.</td>
</tr>
<tr>
<td>SMTP Config</td>
<td>Choose a preconfigured SMTP server to use when sending the email. This also defines who the email will show as being sent from. For more information, see Email Alert section.</td>
</tr>
</tbody>
</table>
6. Workflows

<table>
<thead>
<tr>
<th>Field</th>
<th>Setting to specify</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>Specify the template for email subject.</td>
</tr>
<tr>
<td></td>
<td>The template can contain dynamic values that will be obtained from the crawled</td>
</tr>
<tr>
<td></td>
<td>content (e.g. [cs:PageUrl]).</td>
</tr>
<tr>
<td>Email Body Template</td>
<td>Specify the template for email body.</td>
</tr>
<tr>
<td></td>
<td>The template can contain dynamic values that will be obtained from the crawled</td>
</tr>
<tr>
<td></td>
<td>content (e.g. [cs:PageUrl]).</td>
</tr>
</tbody>
</table>

**TIP:** To get the list of available fields, click the **details** link.

### 6.3.1.2. Migrate Document

This action can be used to copy or move a document between content sources ('source' to 'destination'). Simple migration copies the file and any document properties and is supported by all content source types. Migration action properties specific for different content source types are listed in the table below.

<table>
<thead>
<tr>
<th>Type</th>
<th>As 'source'</th>
<th>As 'destination'</th>
<th>Migration Config Type</th>
<th>Supports structured migration?</th>
<th>Move?</th>
<th>Update source item?</th>
<th>Mark source 'read-only'?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchange</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>File System</td>
<td>Yes</td>
<td>Yes</td>
<td>Custom - File Share</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Google Drive</td>
<td>Yes</td>
<td>Yes</td>
<td>Source (Google Drive account)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>SharePoint</td>
<td>Yes</td>
<td>Yes</td>
<td>Custom - SharePoint Site Collection</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>SQL and other databases</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
**IMPORTANT!** Before you add the Migration action to your workflow, you should configure migration destinations. See [Configuring Destinations for Migration Action](#).

When running the Workflow wizard and having selected Migration as action, you will be prompted to configure related settings.

**To configure migration using Workflow wizard:**

On the **What do you want to do** step, select Migrate Document action. do the following:

1. Specify migration source and folder:

   - Select migration destination under **Which type of repository should the document be migrated to?**. You can add migration destination directly from wizard:

   ![Add Workflow](image)

   - If you created several sources for migration destinations, select one in the under **Where should the document be migrated to?**

   - For Google Drive, you need to specify subfolder to save your files in the **Where in the destination should the files be saved?** field.

2. Configure migration options:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replicate folder structure</td>
<td>If supported by the source system, subfolders will be created in the migration destination to match the relative path in the source. In the case of Exchange this will also include a folder for the mailbox name (I.E: \MigrationDestination\<a href="mailto:User@domain.com">User@domain.com</a>\Inbox\HR).</td>
</tr>
<tr>
<td>Copy or Move the document</td>
<td>Select one of the following:</td>
</tr>
<tr>
<td></td>
<td>- Copy</td>
</tr>
<tr>
<td></td>
<td>- Move</td>
</tr>
</tbody>
</table>
6. Workflows

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark Source as Read-only</td>
<td>The original item can be marked as read only.</td>
</tr>
<tr>
<td>What action should be taken if</td>
<td>Select action to perform:</td>
</tr>
<tr>
<td>the document already exists at</td>
<td>• Replace</td>
</tr>
<tr>
<td>the destination</td>
<td>• Append</td>
</tr>
<tr>
<td>Redact the document</td>
<td>If update of the source item is supported by the source system, then using</td>
</tr>
<tr>
<td></td>
<td>this option will instruct the program to apply the redaction plan to the</td>
</tr>
<tr>
<td></td>
<td>source document after its successful migration.</td>
</tr>
</tbody>
</table>

**NOTE:** This option is not available when performing a move (deleting the original item).

To modify action settings for the certain workflow, select the workflow and use the Advanced UI window. See [Modify Migration action settings](#) for more information.

### 6.3.1.2.1. Modify Migration action settings

To configure or modify Migration action settings using the **Advanced** interface:

1. In administrative web console, navigate to [Workflows](#) and select the workflow you want to configure action for.
2. Click the workflow, then click **Add** next to **Rule Actions**.
3. In the **Add Action** dialog, select the necessary migration action in the **Action Type** list.

There are common and content-specific settings that you need to specify.

#### Common settings

These settings are the same for all supported sources.
### Setting | Description | Comments |
--- | --- | --- |
**Migration Destination** | The root destination to migrate to. | Make sure to define the required destination as Migration Config. |
**Destination Rename Mode** | Specifies what action to take if a file exists at the destination with the same name. | |
- **None** - overwrite the destination file or issue a “duplicate” error. |
- **Append Number** - append a numeric counter as a suffix to the file name (e.g. `document_2.txt`). |
- **Append Date** - append workflow execution timestamp. |
**Maintain Folder Structure** | If selected, subfolders will be created in the migration destination to match the relative path in the source. | Applies if this capability is supported by the source system. For Exchange, the path will also include a folder for the mailbox name (e.g. `\MigrationDestination\User@domain.com\Inbox\HR`). |
### 6. Workflows

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delete Original Item</td>
<td>If selected, the original item will be deleted after it is successfully copied to the destination.</td>
<td>Applies if this capability is supported by the source system.</td>
</tr>
<tr>
<td>Mark Original item as Read Only</td>
<td>If selected, the original item will be marked as <code>read-only</code>.</td>
<td>Applies if this capability is supported by the source system.</td>
</tr>
<tr>
<td>Redaction Plan</td>
<td>If redaction plans have been configured, specify the redaction plan to be applied to the document. See <a href="#">Redaction</a>.</td>
<td>By default, this will be applied to the document at the destination.</td>
</tr>
<tr>
<td>Redact Original</td>
<td>If updating the source item is supported by the source system, then checking this box will cause the redaction plan to be applied to the source document after being successfully migrated.</td>
<td>Note that this option is not available when performing a move (deleting the original item).</td>
</tr>
</tbody>
</table>

**Source-specific settings**

Settings for Google Drive content migration are described below.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Destination Folder</td>
<td>The path to migrate the document to relative to the migration destination.</td>
<td>To migrate to the root folder, leave blank.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Destination example: <code>Folder/SubFolder/SubFolder2</code></td>
</tr>
</tbody>
</table>

Settings for SharePoint content migration are described below.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library/Folder</td>
<td>The library and optional subfolder to migrate the document to in the migration destination.</td>
<td>Only applicable if the source system is also SharePoint.</td>
</tr>
</tbody>
</table>
### 6. Workflows

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dynamic Destination Field Name</strong></td>
<td>Specify a metadata field on the item that can be used to dynamically lookup the migration destination.</td>
<td>Only applicable if the source system is also SharePoint.</td>
</tr>
<tr>
<td><strong>Web Path</strong></td>
<td>The relative web path for the migration of the document. Format should be any of the following:</td>
<td>Only applicable if a SharePoint relative migration is chosen.</td>
</tr>
<tr>
<td></td>
<td>• ~/WebPath — a document found, e.g., at <a href="http://sharepoint/sites/Test/Demo">http://sharepoint/sites/Test/Demo</a> with the relative path</td>
<td></td>
</tr>
<tr>
<td></td>
<td>~/Subsite would attempt to migrate to <a href="http://sharepoint/sites/Test/Demo/subsite">http://sharepoint/sites/Test/Demo/subsite</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• /SiteCollectionPath — a document found, e.g., at <a href="http://sharepoint/sites/Test/Demo">http://sharepoint/sites/Test/Demo</a> with the relative path</td>
<td></td>
</tr>
<tr>
<td></td>
<td>/Subsite would attempt to migrate to <a href="http://sharepoint/sites/Test/Subsite">http://sharepoint/sites/Test/Subsite</a></td>
<td></td>
</tr>
<tr>
<td><strong>List Title</strong></td>
<td>The name of the library at the web path specified to migrate the document to.</td>
<td>Only applicable if a SharePoint relative migration is chosen.</td>
</tr>
<tr>
<td><strong>Fallback - if relative path invalid</strong></td>
<td>Enables/disables falling back to the standard migration destination if the relative path is unavailable. If the relative path does not exist, and the fallback mode is not enabled, then the Workflow will report a failure.</td>
<td>Only applicable if a SharePoint relative migration is chosen.</td>
</tr>
</tbody>
</table>

### 6.3.1.3. Apply Additional Classification

You can instruct the program to apply one or more additional classifications to the processed document. This workflow action is called **Manual Classification** and can be configured via the **Advanced UI** window. See [Advanced Workflow Actions](#) for details.

Alternatively, you can configure a workflow action that permanently removes all existing classifications on a document and disables future auto-classification for it.
To apply additional classification:

In the Add Action dialog, from the Action Type list select Manually Classify under Classification, then configure the necessary terms as described below.

**NOTE:** The terms you select must belong to a single taxonomy / termset.

To remove all classifications:

In the Add Action dialog, from the Action Type list select Remove Classifications under Classification.

To configure terms

1. In the Select Term field, click the tag icon.
2. In the Details dialog, specify filter settings to use when filtering out the documents:
   a. **Taxonomy** - select what classification taxonomy from the existing ones should be used.
   b. **All Terms** - select this option if you want to filter by all terms in the taxonomy. If this option is cleared, then after selecting the necessary taxonomy, you will be presented the list of its terms. Select the one you plan to use for filtering.

   **NOTE:** Multiple selection is not supported: to configure several filter values, you should repeat this procedure for each filter value you need.
   c. **Include Children** - select this option if needed.
3. Finally, click OK to save the settings and close the dialog.

**NOTE:** The additional classification will not trigger other workflows or affect the source item.

6.3.1.4. Advanced Actions for Exchange

In addition to the Email Alert, Migrate Document and additional classification, the following actions are available for the Exchange content source type:
- Delete email
- Move email

To configure these actions, use the Advanced UI dialog window. See [Using Advanced Interface](#) for details on how to invoke it.

### 6.3.1.4.1. Delete Email

This action will remove an email from Exchange mailbox.

![Add Action](image)

Specify the following action parameters:

<table>
<thead>
<tr>
<th>Action parameter</th>
<th>Description</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Delete Mode</strong></td>
<td><strong>Matches the native Microsoft Exchange Delete Modes:</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Soft Delete — Email will be available for recovery from the Deleted Items folder.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Hard Delete — Email will not be available for recovery after deletion.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Move to Deleted items — Email will be moved to Deleted Items folder.</td>
<td></td>
</tr>
<tr>
<td><strong>Suppress Read Receipts</strong></td>
<td>With this option selected, Read receipts will not be sent (if requested) for the item being deleted.</td>
<td>Selected by default.</td>
</tr>
</tbody>
</table>

### 6.3.1.4.2. Move Email

This action will move an email to the specified folder within the same mailbox.

Specify the following action parameters:
### Action parameter | Description | Comments
--- | --- | ---
**Target Folder Name** | The name of the folder the move the email to. | For subfolders, only include the subfolder name (not the full path).

**Parent Folder Name** | If the target folder name is not unique, specify the parent folder name — to ensure the correct folder is used. | Optional.

---

**6.3.1.5. Advanced Actions for File System**

In addition to the Email Alert, Migrate Document and additional classification, the following actions are available for the File system content source type:

- **Update Permissions** — this action updates the file system permissions for the classified document. See [Update Permissions](#) for details.

- **Apply MIP Label, Remove MIP Label** — these actions, respectively, apply and remove sensitivity label to a document stored on a file system, using Microsoft Information Protection (MIP). This helps to automate protection policies application.

To configure actions for file systems using the Advanced interface:

1. In administrative web console, navigate to Workflows and select the workflow you need.
2. Click the workflow, then click Add next to Rule Actions.
3. In the Add Action dialog, select the action you need from the File System section in the Action Type list.

**NOTE:** To apply or remove MIP label, a MIP application configuration must be specified. See [Configure Infrastructure](#) for more information.
6.3.1.6. Advanced Actions for SharePoint

In addition to the Email Alert, Migrate Document and additional classification, the following actions are available for the SharePoint content source type:

- **Migrate Document** including copy and move operations
- **Document property field (metadata) update**, including:
  - **Send fixed value, send crawled value** — these actions apply new metadata value entered by user or retrieved from the related NDC database field, respectively.
  - **Send classification value** — this action writes classification metadata (Taxonomy) into the selected property field (Field Name). If multiple classification values are applied, they will be written using delimiters.
  - **Write O365 Label, Remove O365 Label** — use these actions to write or remove Office 365 retention label as document metadata. These labels are typically used to automatically apply data protection policies to your documents.

**NOTE:** These actions require Microsoft Office 365 retention labels to be configured. See this [Microsoft article](#) for details.

- **Filtered Targeted Meta Update** — this advanced action can be used to update a SharePoint property based on rules embedded in the taxonomy clues. Enter the document property to update in the **Update Field**, then select the required **Taxonomy** and enter **Match Field**, i.e. the field name/clue to match on.

To configure actions for SharePoint documents using the Advanced interface:
1. In administrative web console, navigate to **Workflows** and select the workflow you want to configure action for.

2. Click the workflow, then click **Add** next to **Rule Actions**.

3. In the **Add Action** dialog, select the action you need from the **SharePoint** section in the **Action Type** list.

![Add Action dialog](Image)

**6.3.1.6.1. SharePoint Content Type Hubs**

SharePoint 2010+ supports **Enterprise Content Types** allowing **Content Types** to be defined on a Publishing SharePoint site with one or more secondary sites consuming the Enterprise Content Types.

Once Netwrix Data Classification for SharePoint is installed on the SharePoint Farm, it is possible to define SharePoint workflow actions at the SharePoint Content Type Hub site. Any actions of type **Content Type Update** may be run on the site collection itself however they may also be run on consuming SharePoint Site collections.

![Netwrix Data Classification 5.5.1](Image)

To configure a Workflow to run against all sites that consume a Content Type Hub please follow the below steps:

1. Navigate to **Workflows** → **Configs** → **Content Type Hubs**

2. Select **Add**

3. Enter the connection details for the **Content Type Hub Site Collection**
4. Once added, navigate back to the main Workflows screen, and select the newly added group from the Workflow Groups grid.

5. Finally, select Add and create the Workflow as normal.

### 6.3.2. Plugins for Additional Actions

In addition to the common workflow actions provided out-of-the-box, you can set up additional actions using the plugins. Either use sample plugins from the vendor, or create your own custom plugins. Plugins should be stored in the dedicated folder, under `C:\Program Files\ConceptSearching\Plugins\`.

The following sample plugins (implemented as DLLs) can be provided upon request:

- FTP Migration action
- Http Save Files action
- Twitter action
- SQL Lookup

To search for the plugins within default location, go to the Plugins tab and click Detect Plugins.

Click the Enable link to enable selected plugins.

To modify workflow action implemented by a plugin, go to the Configs tab and click Action Configs on the left.

### 6.4. Workflow Operations Log

When workflow actions are performed, the corresponding operations are logged to the web-based log file. Click the Logs tab to view the corresponding audit trails.

Here you can change the display period or the number of logs displayed, sort the list or copy its content, or clear the logs you do not need.
6.5. Workflow Plugins

A range of Workflow actions are provided with the product, but the product can also be extended by writing additional actions using the plugin interfaces.

Plugins are implemented as DLLs and are placed in the plugins folder, which is typically located here: C:\Program Files\ConceptSearching\Plugins\

The following sample plugins are provided with the product (complete with code):

- FTP Migration action
- Http Save Files action
- Twitter action
- SQL Lookup

Click the Detect New Plugins button to search the plugins folder for new plugins.
Click the Enable link to enable selected plugins.
7. Administrative Tasks

This section describes the operations that you can perform when administering your Netwrix Data Classification using the management console, in particular:

- Configuration
- Index Maintenance
- Configuration Backup
- Review Dashboards

7.1. Index Maintenance

**NOTE:** Only available for ‘Superusers’.

You may need to reprocess content or even clean the environment on a large scale — for example, after a large amount of content has been deleted, or after configuring a DQS environment. In such scenarios, index should also be maintained — to ensure data consistency. To automate maintenance operations, you can use a built-in tool named Cleaner.

**To launch the Cleaner tool**

2. Navigate to Config → Settings and click Run Cleaner.
3. Then follow the steps of Index Maintenance wizard.
See next:

- **Step 1: Maintenance Operation**
- **Step 2: Maintenance Options**
- **Step 3: Summary**
- **Step 4: Process**

### 7.1.1. Step 1: Maintenance Operation

Select the operation you want to perform:

- **Rebuild Index**—All content processing results (text/metadata) will be retained, but the search index will be truncated. Then the program will re-do all indexing/classification (during that process, search results will be unavailable). Optionally you can choose to Shrink - this will rebuild the Text.cse file removing any fragmentation. Shrink will require sufficient disk space to process (up to the existing size of Text.cse)

- **Re-Collect Index**—The search index will be cleaned (all documents from the source will be removed from it). Then the program will re-crawl all configured sources and update the search index (during that process, search results will be unavailable).

**NOTE:** This option is recommended after setting up DQS configuration.

- **Delete Index**—Delete all content from both the search index and the NDC SQL database.

### Index Maintenance

- **Rebuild Index**
  - Maintain all source/content and rebuild the core search index
- **Re-Collect Index**
  - Maintain all sources, delete all content, and rebuild the core search index
- **Delete Index**
  - Delete all sources and reset the application

### 7.1.2. Step 2: Maintenance Options

Specify options for the operation you have selected.

<table>
<thead>
<tr>
<th>Operation selected</th>
<th>Available options</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rebuild Index</td>
<td>Shrink the &quot;text.cse&quot; file?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>◦ Shrink</td>
<td>Selecting <strong>Shrink</strong> will rebuild the Text.cse file, removing any fragmentation. <strong>Shrink</strong> will require</td>
</tr>
</tbody>
</table>
### 7.1.3. Step 3: Summary

Review the selected operation (action) and its options you have specified.

Clicking **Next** will confirm and start the maintenance operation.

### 7.1.4. Step 4: Process

Finally, wait for the selected maintenance operation to complete. Until then, search results will be unavailable.

### 7.2. Configuration Options

The **Config** administration area provides a web based console for altering global system configuration settings. The default screen shows the most commonly amended settings.
The most heavily used settings are displayed by default. Some configuration options are hidden and can be shown by selecting the **Advanced Settings** screwdriver. Optionally, users can choose to always see advanced settings as part of their user preferences. See **Security (Users)** for more information.
7.2.1. Core Configuration

Each configuration option has an associated “I” which describes the nature of the setting. Selecting the Details tab provides a complete list of the Config settings – as well as an indication of the values that have been changed from the default setting.

You can also:

- **Reset QS Cache**—Force the QS caches to be reset.
- **Run Product Tour**—Runs a product tour, taking you around the key areas of the product.

7.2.2. Licensing

The licenses that are loaded into the product define what functionality is available. This is broken into:

- **Sources**—Sources that are available to be added / crawled
- **Tagging Write Back**—Sources that are available to be have classifications written back to the repository
- **Features**—Redaction, Custom Reporting, Clue Building Reports, and Workflow capabilities

The default licensing display provides a summary of the current license state. Select **Add License** to load / update a license. You can also view and manage the available license by selecting **Licenses** from the side menu.
7.2.3. Metadata Configuration

This section contains information on how to configure metadata of your documents. Review the following for additional information:

- Document Metadata Fields
- Metadata Field Mappings
- Metadata Value Mappings

Document Metadata Fields

This list specifies which internally generated fields are to be used:

<table>
<thead>
<tr>
<th>Field</th>
<th>Ignore</th>
<th>Edit</th>
<th>Delete</th>
</tr>
</thead>
<tbody>
<tr>
<td>_isCurrentVersion</td>
<td></td>
<td>Edit</td>
<td>Delete</td>
</tr>
<tr>
<td>_ModerationStatus</td>
<td></td>
<td>Edit</td>
<td>Delete</td>
</tr>
<tr>
<td>_reportinggallerymetadataid</td>
<td></td>
<td>Edit</td>
<td>Delete</td>
</tr>
<tr>
<td>_reportinggallerytemplateid</td>
<td></td>
<td>Edit</td>
<td>Delete</td>
</tr>
<tr>
<td>Abstract</td>
<td></td>
<td>Edit</td>
<td>Delete</td>
</tr>
<tr>
<td>allowslistpolicy</td>
<td></td>
<td>Edit</td>
<td>Delete</td>
</tr>
<tr>
<td>AnonymousViewMask</td>
<td></td>
<td>Edit</td>
<td>Delete</td>
</tr>
<tr>
<td>Attachments</td>
<td></td>
<td>Edit</td>
<td>Delete</td>
</tr>
<tr>
<td>BaseType</td>
<td></td>
<td>Edit</td>
<td>Delete</td>
</tr>
</tbody>
</table>

Metadata Field Mappings

This table allows additional metadata fields to be generated by mapping an already existing field name to a new name.

<table>
<thead>
<tr>
<th>Source Field Name</th>
<th>Target Field Name</th>
<th>Edit</th>
<th>Delete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>PageTitle</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


For example, if we create an entry with Source=Author and Target=Publisher then a document with this metadata:

“Author: John Challis;”

Will generate an index with this metadata:

“Author: John Challis; Publisher: John Challis;”

This facility can be useful when you need to align metadata field names across a variety of sources and/or document types.

*Metadata Value Mappings*

This list allows metadata values to be mapped from a source value to a new target value.

For example, if we create an entry for the field “Modified By”, with Source=“Cheryl Tweedy” and Target=“Cheryl Cole”, then a document with this metadata:

“Modified By: Cheryl Tweedy;”

Will generate an index with this metadata:

“Modified By: Cheryl Cole;”

This facility can be useful when you need to align metadata field values for example when employees change their name or are replaced by different people.
7.2.4. Email Configuration

This section contains information on how to configure email servers for external communication, including configuring email groups and health service notifications. Review the following for additional information:

- Email Servers
- Email Groups
- Health Service Notifications

Email Servers

Email servers can be configured to enable external communication. For instance when the health service identifies an issue.

Servers can be amended post configuration by selecting **Edit**, or, new SMTP servers can be added by selecting **Add Email Server Configuration**.

The SMTP details should be entered based on the values provided by your network team. Each configuration supports both SSL enabled SMTP servers, and those without SSL enabled.

It is also possible to supply a test email address which will be used to test the configuration settings.
Email Groups

Email groups are used to define a logical group of people to email, essentially – a mailing list.

Each email group is linked to an SMTP server, so, before configuring an email group, you must configure your Email Servers.

To add a new group, select Add Email Server Group, or select Edit on each row to configure the group members.

Each group can have one or more members, and can be assigned a friendly name, which will be displayed when selecting an email group.

Health Service Notifications
**Health Service Notifications** can be configured to email a specific group of people when something goes wrong within the product. Each notification configuration is linked to an email group, so, before configuring notifications, you must configure your Email Groups.

To add a new notification configuration select **Add Notification Configuration**, or select **Edit** on each row to change the configuration.

Notifications can be set to trigger on warnings, or just on errors – by default problems of any level will be reported.

The **Daily Summary** can also be disabled / enabled, this functionality sends out a summary email of outstanding problems each morning.

---

### 7.2.5. Text Handling

This section contains information on how to configure text processing. Review the following for additional information:

- [Best Bets](#)
- [Content Type Extension Mapping](#)
- [Content Type Extraction Methods](#)
- [Language Detection](#)
- [No Stem](#)
- [OCR Language Mapping](#)
Best Bets

Sometimes an application may wish to push selected documents to the top of a hitlist for specific queries. This may be implemented by specifying **Best Bets** for specific query text.

First, enter the search term that you wish to match and then click the **Add** button.

Next, click on the term, and specify one or more URLs that should appear at the top of the hit list.

**Content Type Extension Mapping**

Sometimes an organization may wish to process certain file types as a different content type. The primary use case for this is internal content types that map to a content type already understood / identified.

In this case the example has a .rpt file being treated as a text file, as such the file will be copied to a temporary location as a .txt file and processed as if it were any other text file.

**Content Type Extraction Methods**

The **Content Type Extraction** methods describes how documents will be handled by the APIs and the core services. A number of built-in processing methods are available, where there is no available method the processing will default to running through standard Microsoft Search iFilter processing.

The methods can be easily altered by clicking **Edit** and then selecting the preferred processing method. It is also possible to specify that an iFilter should be utilised if the primary method fails to extract text from the document – the backup method will be used if the extraction fails to find more than 5 characters of text.
If you have updated the extraction method we recommend re-processing any documents that have already been processed to ensure consistency. Selecting **Re-index** from the grid for the affected content type will re-process the necessary records.

<table>
<thead>
<tr>
<th>Content Type</th>
<th>Default Extension</th>
<th>Extraction Method</th>
<th>Use iFilter as Backup</th>
<th>Re-Index</th>
<th>Edit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adobe Photoshop</td>
<td>.psd</td>
<td>iFilter</td>
<td></td>
<td>Re-Index</td>
<td>Edit</td>
</tr>
<tr>
<td>AIFF</td>
<td>.aiff</td>
<td>iFilter</td>
<td></td>
<td>Re-Index</td>
<td>Edit</td>
</tr>
<tr>
<td>Archive</td>
<td>.zip</td>
<td>iFilter</td>
<td></td>
<td>Re-Index</td>
<td>Edit</td>
</tr>
<tr>
<td>Bitmap</td>
<td>.bmp</td>
<td>iFilter</td>
<td></td>
<td>Re-Index</td>
<td>Edit</td>
</tr>
<tr>
<td>CAD</td>
<td>.dwg</td>
<td>Aspose</td>
<td></td>
<td>Re-Index</td>
<td>Edit</td>
</tr>
<tr>
<td>Compiled HTML</td>
<td>.chm</td>
<td>iFilter</td>
<td></td>
<td>Re-Index</td>
<td>Edit</td>
</tr>
<tr>
<td>DICOM</td>
<td>.dcm</td>
<td>iFilter</td>
<td></td>
<td>Re-Index</td>
<td>Edit</td>
</tr>
</tbody>
</table>

**Language Detection**

The language detection list specifies which languages will be considered for auto-detection.

If a language is excluded then it cannot be used to identify the language of a document and it will be removed from the language options in Taxonomy Manager.

**No Stem**
The **No Stem** list offers the ability to disable language stemming for a particular word or phrase, this supports the ability to always apply a phrasematch when a particular term is used as either a clue – or a search term.

![No Stem](image)

**OCR Language Mapping**

The **OCR** language mapping configuration screen can be used if you wish to OCR non-English images via Tesseract. File paths (including parts of paths) can be mapped to specific Tesseract language packs.

![OCR Language Mapping](image)

**Synonyms**

Often it is important to submit a query and have synonyms automatically included. A generic set of synonyms may be configured by using the **Synonyms** form.

![Synonyms](image)

**Text Patterns**
Many HTML web pages contain navigation information and other extraneous information that is the same for all pages and/or not relevant to the individual page content. If all of the text is indexed from these HTML pages then this can lead to unwanted search results where a match is made, for example, to an entry in a standard page navigation area.

The Text Patterns feature is provided to assist with the cleanup of HTML documents. TextPatterns can also be used to index terms that would normally be discarded.

The StartTag and EndTag values are case sensitive strings used to identify the content to be managed, the content is then managed based on the filter type.

There are three tag types that can be used to assist in the cleanup:

- **FILTER**—Extracts a subset of the HTML page, prior to extracting the plain text. Only a single section will be extracted for each TextFilter processed.

- **DELETE**—Deletes sections of the HTML page, prior to extracting the plain text.

- **INDEX TERM (EndTag ignored)**—Create index terms that would otherwise not be formed. For example the term “E.ON” is a useful one for people interested in energy companies. However, this term would not normally be created because a full stop normally acts as a term separator. However, if we create an INDEX TERM for this pattern then it will be detected and indexed as required.

### 7.2.6. Redaction

This section contains information on configuring redaction plans and entities. Review the following for additional information:

- **Redaction Plans**

- **Redaction Entities**

**Redaction Plans**

Redaction plans can be used as an optional migration step to remove specific entities from supported content types. During the migration of a document a migration plan will remove the following entity types (depending on the configuration):

- **NLP Entities**—Items identified by the NLP entity extraction, such as names or locations

- **Regex Entities**—Items identified by the Regex classification clues, such as credit card numbers or
social security numbers

- Specific clues can be skipped as part of a redaction plan by specifying **Excluded Clues**, such as: “VISA” or “SSN” (matched to the term name)

- **Custom Entities**—Any custom words or phrases associated with the plan.

Masking based redaction will ensure that a specified number of start / end characters will be retained from each redacted value.

![Redaction Plans](image)

**Redaction Entities**

Redaction entities can be used to specify any custom words or phrases that should be removed by a redaction plan.

![Entities](image)

### 7.2.7. Additional Configuration Settings

This section contains information on additional configuration settings specific to different source types.

- **AD Domains Excluded**
- **Attachments Excluded**
- **No Index**
- **Proxy Server**
- **Suspend Services (Scheduler)**

**AD Domains Excluded**
The **AD Domains Excluded** list is used to disable Active Directory expansion for certain domain names. This is useful in a multi-Domain forest, where the Netwrix Data Classification server does not have access to all domains within the forest.

**Attachments Excluded**

When indexing files from that potentially contain attachments (SharePoint List Items) the list of file locations that will be ignored is defined by the **Attachments Excluded** list. The definitions in this list may be viewed and modified via the Attachments Excluded form:

Any file with a path that matches one of these patterns will be ignored. Wildcards may be used anywhere in the pattern definition, with:

- The asterisk character (*) matching any sequence of characters
- The question mark character (?) matching any single character

**No Index**

Sometimes an application may wish to remove selected documents from all search results. This may be implemented by specifying **No Index** entries.
Any number of URLs (or Filenames) may be entered and none of these will ever appear in search results. Wildcards may be used anywhere in the pattern definition, with:

- The asterisk character (*) matching any sequence of characters
- The Question mark character (?) matching any single character

**Proxy Server**

The Proxy Server form may be used to define a proxy server to be used when crawling websites, the proxy server is not used for SharePoint crawling.

Set Bypass Local to Yes to bypass the proxy server for local addresses (localhost etc).

Any other exclusions that should not go through the proxy server should be defined in the Exceptions list.

**Suspend Services (Scheduler)**

All Netwrix Data Classification services run as Windows services. They are responsible for building the search index and classifying documents against the registered taxonomies.

It can be useful to suspend these services from running so that they do not impact query performance during the peak hours of the working day. Sometimes it may be useful to suspend these services for some lower priority sources but have them continue to process higher priority sources.
Service suspensions can be configured in the following ways:

- **Source**—Which source types the suspension is in place for: all source types, specific source types (SharePoint, Web etc) or specifically against Re-Indexing operations.
- **Service**—Which services are affected by the suspension: All Services, or, a choice of: NDC Collector, NDC Indexer, NDC Classifier.
- **Day/Time**—Allows the configuration of which days and times the suspension will be in place.

### 7.2.8. Configuration Backup

**NOTE:** Only available for ‘Superusers’

The **Backup** utility allows for the migration of complex Netwrix Data Classification instance configurations. This allows a user to safely design and test a concept searching configuration within a development environment and then copy the configuration, or specific parts of the configuration, to a different environment (I.E production).

The tool supports text replacement to allow user defined URL’s to be replaced by the equivalent destination URL. The following configuration options are available for import / export:
- Source Registrations
- SharePoint Termset Registrations
- Workflow Configurations
- Core Configuration Options:
  - Files Excluded
  - Files Included
  - Mapped Metadata Fields
  - Mapped Metadata Values
  - Supported Languages
  - Pages Excluded
  - Pages Included
  - SharePoint Excluded
  - Text Patterns

To create a backup simply select **Create Backup** and select the elements that you wish to include. The backup password will be required if you export a backup to XML and re-import to a different environment.

Upon import any items that already exist will be skipped.
### 7.3. Review Dashboards

The **Dashboard** administration area provides a selection of tools to review application health.

The default screen shows a high level overview of service statistics. The last active times of each of the core windows services are shown, with inactive services shown in red. Selecting the "i" icon next to each date will identify the name of the active server as well as batch processing statistics providing an indication of document processing throughput. The following statistics are available for each thread type:

- **Processing Time**—The weighted average time taken for each thread (total batch time / number of documents processed)

- **Real Execution Time**—The actual execution time of each thread (average of each threads run time)

Statistics shown on the **Dashboard** screen are cached and updated regularly by the Collector service. If the values are not being updated please ensure that the Collector service is running.

New content will be shown as awaiting collection, and progress through to fully processed once it has been classified.

Content that has failed to process fully will be indicated under the "Exceptions" section, with the following meanings:

- **Collection Errors**—Items that failed to process during collection (typically due to an error from the source system)

- **Text Extraction Errors**—Items that failed text extraction (either partially or fully)—this will typically mean that the full text for the affected documents will not be available

- **Collection Exclusions**—Items that have been excluded due to the specified configuration (such as Sources → SharePoint → Exclusions)

- **Files Skipped**—File share items that have been ignored due to the "Files Included" or "Files Excluded" configuration (Sources → File)

- **Deleted Automatically**—Items that have been detected as removed from the source system
Deleted Manually—Items removed manually by an end-user via the administration console

7.3.1. System Health

The health service provides a traffic light based reporting system. Colour-coded traffic lights will appear in the top menu bar when issues are detected. The traffic lights provide a quick link to this page to display more detailed information.

You will then see the list of reported issues, with the ability to view a detailed description of the problem and suggested resolution steps.

It is also possible to configure notifications of system issues, along with daily reports of outstanding system issues. Please see the Health Service Notifications configuration for more details.

7.3.2. Netwrix Data Classification Service Viewer

From the Netwrix Data Classification Service Viewer it is possible to view a live stream of the current work being processed by the Windows Services. As the services progress each document the display will change. Once all work is complete "Idle..." will be displayed.

This functionality may not work in older browsers. In this case the "on server" application Netwrix Data Classification Service Viewer should be used.
8. Reporting Capabilities

The **Reports** administration area helps a user extract a wealth of information from the Netwrix Data Classification index. The main dashboard has three high level graphs highlighting the current state of processing:

- **Document Progress**—A graphical display of the main stats display, once processing is complete documents will be allocated to either Fully Processed or Errors
- **Index Size**—Shows the percentage of each source type being processed: Files, SharePoint, SQL and Web sources
- **Classification Coverage**—Shows the percentage of classified content, broken down by type, and the percentage of content that has not received any auto-classifications

The **Content Distribution** report highlights areas of classification overlap.

It is possible to filter and refine data presentation to look for the areas that contain the largest amount of documents tagged with a particular term, or to only review specific content.

Review the following for additional information:

- [Content Distribution](#)
- [Review Built-in Reports](#)
- [Types of Reports](#)
8.1. Content Distribution

The Content Distribution treemap allows you to interrogate your data in two different ways:

- **Taxonomy Grouping**—When grouped by taxonomy the treemap will highlight the sources with the largest numbers of documents tagged to the selected taxonomies / terms. An example use case is when detecting sensitive documents—in this case the treemap will show the highest risk sources. Clicking on an area will drill into the term / taxonomy to provide a clear view of the affected sources. Optionally "safe" sources, such as quarantine locations, can be excluded.

- **Source Grouping**—When grouped by source the treemap will show the level of content either untagged or tagged within the taxonomy. The treemap will display the top level terms for the selected taxonomy with the counts including any document that is tagged to either the top level term or one of its descendants. Documents can be tagged to one or more terms so the number associated with each top level term may exceed the number of documents in the source.

It is possible to filter and refine this display, either selecting specific sources / source-groups or excluding specific sources / source-groups.

8.2. Review Built-in Reports

Reports can be found under the Reports tab.

There are a number of reports provided that can be run in browser, as well as exported to excel, these are described below:

- **Classification Coverage**—Provides a list of documents that have been tagged with X or fewer classifications. Assists in locating documents that have a low number of auto classifications and highlights the nearest missed classification. Supports filtering by URL and source group.
- **Classification Misses**—Reports on documents that almost reached the threshold for classification, but ‘missed’ being classified by 20% or less. Supports filtering by URL and source group.

- **Clue Counts**—Provides a report of the number of clues per term, also includes a count of regular expression clues.

- **Clue Coverage**—Provides a report on the usage of clues within classification tagging. Assists in highlighting clues that are not aiding the classification process, or clues that are too vague. Supports filtering by URL and source group.

- **Document Tagging**—Provides a report on the manual and automatically assigned document classifications. Supports filtering by URL and source group.

- **Duplicate Detection**—Provides a list of documents that are considered “duplicates” within the index, using checksum matching. Supports filtering by URL and source group.

- **Failed Write Classifications**—Provides a list of documents in the core index that failed to have their classifications written to the source system (such as SharePoint Managed Metadata Columns). Supports filtering by URL and source group.

- **Files Skipped**—Provides a list of documents that have been excluded from processing because they were not explicitly included, or were specifically excluded. See Files Included and Files Excluded for more information on file inclusion / exclusion. Supports filtering by URL.

- **iFilters Detected**—Provides a list of detected iFilters per server. iFilters are the Microsoft standard for implementing text extraction from binary files. They are used by many search engines (including Microsoft Search) to obtain the plain text required to build a search index. Supports filtering by server.

- **Index Analysis**—Provides the ability to manually queue items for background index analysis, initially scoped to assist in identifying fuzzy matched duplicate documents.

- **Manual Tagging**—Provides a report on the manual and automatically assigned document classifications—filtered specifically to manually classified documents. Supports filtering by URL and source group.

- **Near Duplicate Detection**—Details near duplicate documents across the index. Near duplicates are detected as a background process, to enable the background processing simply enable the option ‘Near Duplicate Detection’ within the NDC Indexer Settings and rebuild the necessary sources. See [Core Configuration](#) for the configuration details. Supports filtering by URL, source group and excluding content types (comma delimited list of content types such as: “css,pdf”).

- **Page Statuses**—Provides a list of documents at a given status within the index. Supports filtering by URL and source group.

- **Term Cloud**—Displays the top 50 key terms/phrases across the index, selecting a term expands the cloud into the related terms.

- **Term History**—Displays a history of changes made to a taxonomy (clues added/deleted etc). Supports filtering by term name.

- **Text Extraction Failures**— Provides a list of documents in the core index that failed text extraction (granular iFilter error codes). Supports filtering by URL, title and source group.
- **Term Links**—Provides a list of links to a specified term (Metadata clues, Term Boosts and Required Term links)—useful when retiring taxonomy nodes to avoid invalid links to the term you wish to remove.

Review the **Near Duplicate Detection** report as an example:

![Near Duplicate Detection](image)

Certain document specific reports can be exported along with any associated document metadata.

![Export](image)

To export specific metadata:

1. First choose one of full export options (CSV/XLSX)
2. Then, tick the **Include Metadata** checkbox
3. Enter the metadata values you wish to export below by starting to type in the text input
4. Click **Export**
8.3. Types of Reports

This section contains description of all types of reports available in Netwrix Data Classification. Review the following for additional information:

- **Auto Classification Review**
- **Queued Reports**
- **Custom Reports**

**Auto Classification Review**

Provides a list of documents tagged with a particular term or terms (using either an AND or OR operator). Optionally incorporates granular information on how documents are being tagged down to the clue level – allowing simple review of the classification configuration. Supports filtering by URL and source group.

![Auto Classification Review](image)

**Queued Reports**

When large search exports are run the report may take some time to compile, in this instance the background processes create the report and make it available for download via the Queued Reports dashboard. Reports can be deleted prior to, or after, processing as well as downloaded as many times as necessary.
**Custom Reports**

While there are a number of reports included in the product by default, it is also expected that specific business needs may arise that require reporting not covered by the default reports.

With this in mind it is also possible to create custom report **Plugins**. Once the custom report plugin is deployed the report will appear in the main reports list with the built-in reports. A sample plugin can be provided which shows a simple example that incorporates:

- Custom Parameters
- Custom Filters
- Report Sorting
- Paging

The application communicates normally with any one of the servers running the administration Web Services. Each server will automatically communicate with the other servers in the cluster to assemble and combine the required search results.