

Netwrix Data Classification User Guide

Version: 5.5.2
3/10/2020



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

2.7.1. Applying DQS Mode	45
2.8. Configure IT Infrastructure	49
2.8.1. Configure Microsoft Exchange for Crawling and Classification	50
2.8.2. Configure NFS File Share for Crawling	53
2.8.3. Configure G Suite for Crawling	53
2.9. Initial Product Configuration	56
2.9.1. Select Processing Mode	56
2.9.2. Processing Settings	57
2.9.3. Add Taxonomy	57
2.9.4. Review Your Configuration	58
3. Security (Users)	59
3.1. Secure Netwrix Data Classification	59
3.2. User Management	61
3.3. Password Manager	64
3.4. Web Service Security	65
4. Content Sources	66
4.1. Add a Content Source	66
4.1.1. Database	67
4.1.2. Exchange Mailbox	68
4.1.3. Exchange Server	69
4.1.4. File System	70
4.1.4.1. Add Folder source	70
4.1.4.2. Add Files source	72
4.1.5. Add Google Drive Source	73
4.1.6. Outlook Mail Archive	75
4.1.7. SharePoint	75
4.1.8. SharePoint Online	76
4.2. Narrow Data Collection Scope	77
4.3. Use Tagging (optional)	77
4.4. Manage Sources and Control Data Processing	78

4.4.1. Modify Source Settings	79
4.4.2. Set up granular processing and tagging for Database	80
4.4.3. Set up exclusions and tagging for Exchange	85
4.4.4. Set up filters and tagging for File System	85
4.4.4.1. Configure Inclusions	85
4.4.4.2. Configure Exclusions	86
4.4.4.3. Configure Tagging	87
4.4.5. Set up exclusions and tagging for Google Drive	89
4.4.6. Set up processing options for SharePoint	90
4.4.7. Set up processing options for SharePoint Online Tenancy	95
4.5. View Results	96
4.5.1. Data Processing Statistics	96
4.5.2. Content Crawling and Classification Results	96
5. Taxonomies	99
5.1. What are Taxonomies?	99
5.2. Built-in Taxonomies Overview	99
5.2.1. Core Taxonomies	100
5.2.2. Derived Taxonomies	102
5.3. Taxonomy Settings	104
5.3.1. Taxonomy Settings Levels	105
5.3.2. Labels	107
5.3.2.1. SharePoint Labels	108
5.3.2.2. O365 Labels	108
5.3.2.3. Help	108
5.4. Add a Taxonomy	108
5.5. Manage Taxonomies	109
5.5.1. Managing Term Sets	114
5.5.2. Multi-User Environments	114
5.6. Search and Filter Taxonomies	115
5.7. Classification Rules (Clues)	118

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

6.2.2.1. Specifying Rule Conditions	157
6.2.2.2. Specifying Rule Actions	159
6.2.2.3. Other Rule Settings	160
6.2.2.4. Specifying Workflow Conditions	160
6.2.3. Edit Workflow settings	162
6.2.4. Delete Workflow	163
6.3. Workflow Actions	164
6.3.1. Available Actions	164
6.3.1.1. Email Alert	165
6.3.1.2. Migrate Document	168
6.3.1.3. Apply Additional Classification	173
6.3.1.4. Advanced Actions for Exchange	174
6.3.1.5. Advanced Actions for File System	176
6.3.1.6. Advanced Actions for SharePoint	177
6.3.2. Plugins for Additional Actions	179
6.4. Workflow Operations Log	179
6.5. Workflow Plugins	180
7. Administrative Tasks	181
7.1. Index Maintenance	181
7.1.1. Step 1: Maintenance Operation	182
7.1.2. Step 2: Maintenance Options	182
7.1.3. Step 3: Summary	183
7.1.4. Step 4: Process	183
7.2. Configuration Options	183
7.2.1. Core Configuration	185
7.2.2. Licensing	185
7.2.3. Metadata Configuration	186
7.2.4. Email Configuration	188
7.2.5. Text Handling	190
7.2.6. Redaction	194

7.2.7. Additional Configuration Settings	195
7.2.8. Configuration Backup	198
7.3. Review Dashboards	200
7.3.1. System Health	201
7.3.2. Netwrix Data Classification Service Viewer	201
8. Reporting Capabilities	202
8.1. Content Distribution	203
8.2. Review Built-in Reports	203
8.3. Types of Reports	206

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-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.

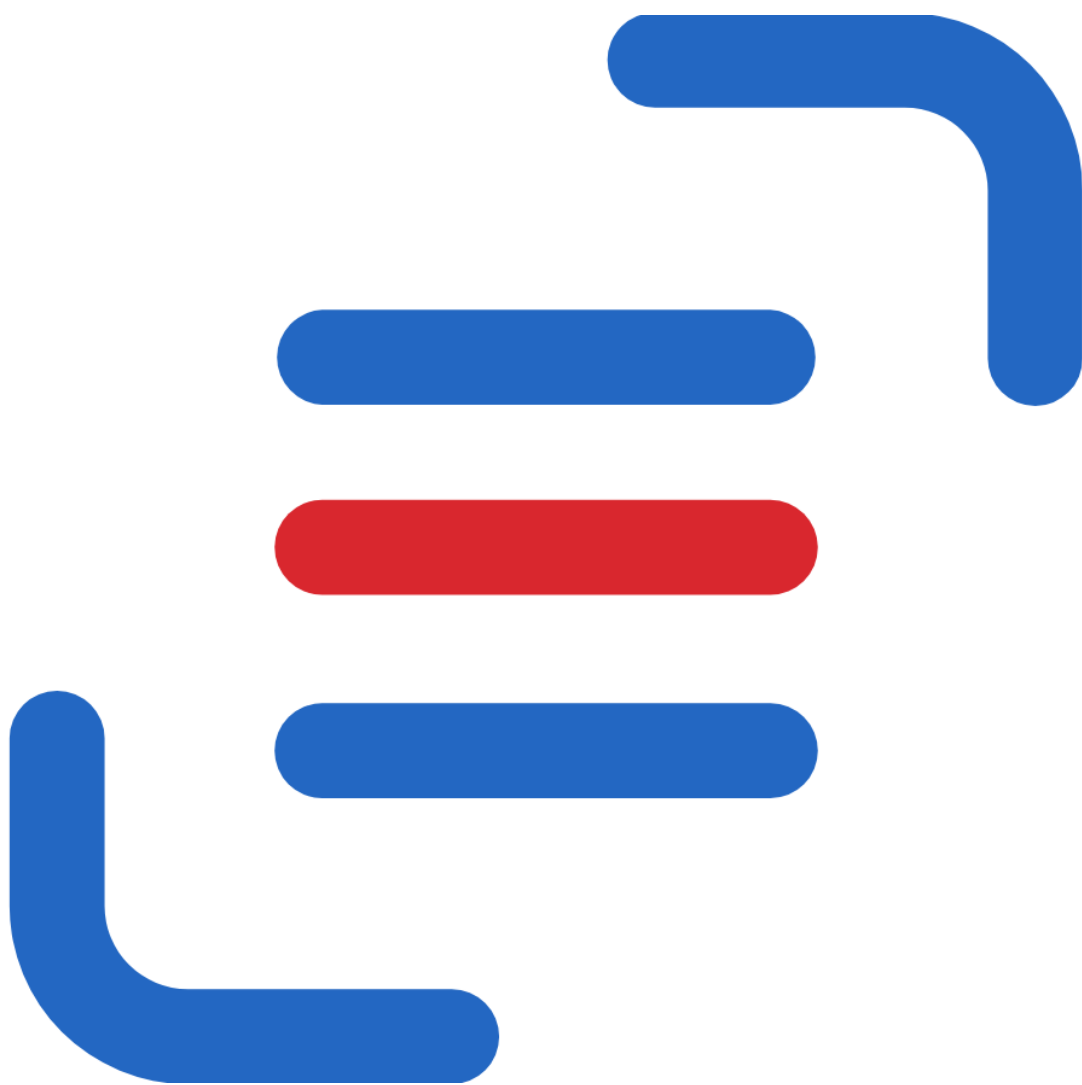
© 2019 Netwrix Corporation.

All rights reserved.

2. Deployment

Netwrix Data Classification User Guide

Version: 5.5.2



3/10/2020

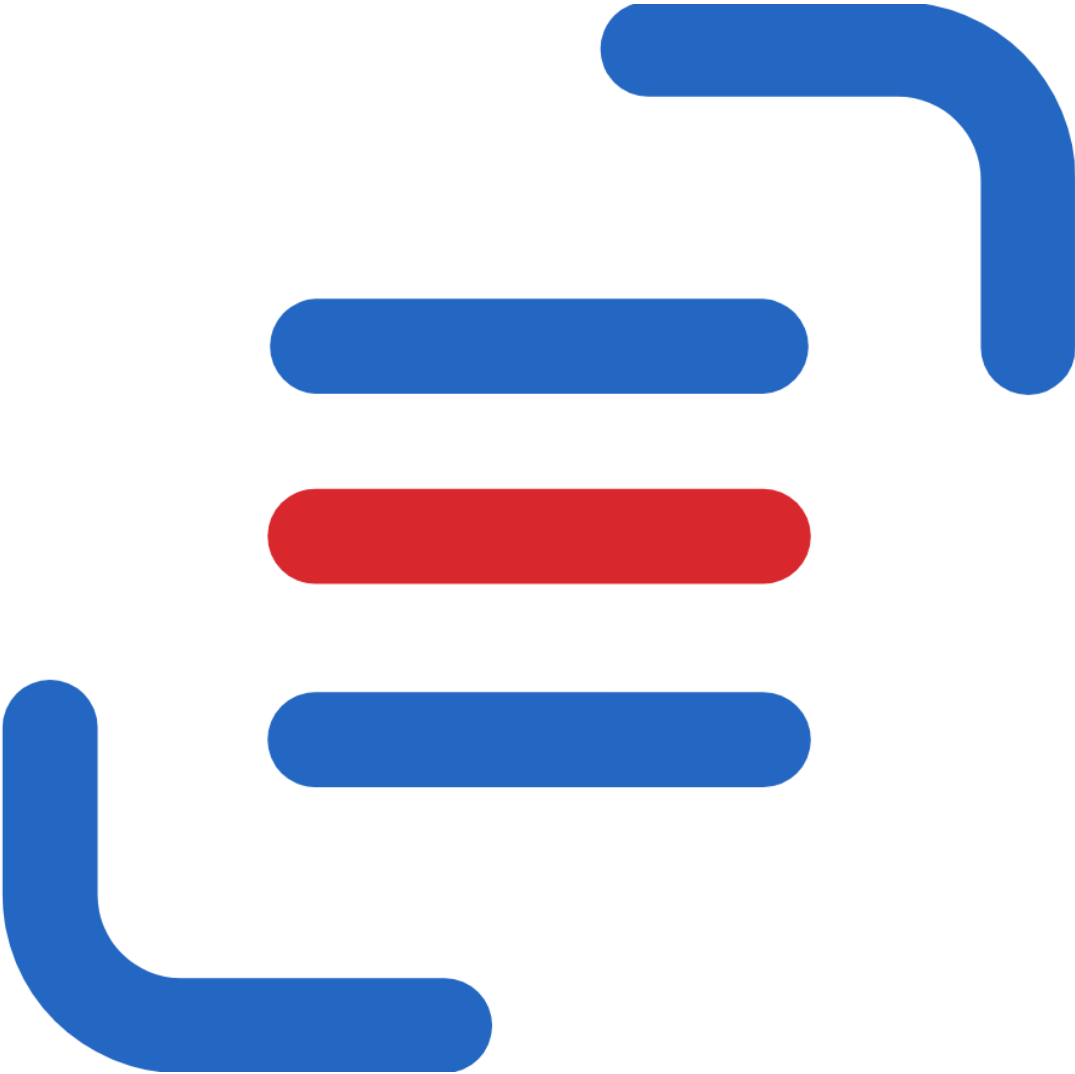


Table of Contents

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

2.7.1. Applying DQS Mode	45
2.8. Configure IT Infrastructure	49
2.8.1. Configure Microsoft Exchange for Crawling and Classification	50
2.8.2. Configure NFS File Share for Crawling	53
2.8.3. Configure G Suite for Crawling	53
2.9. Initial Product Configuration	56
2.9.1. Select Processing Mode	56
2.9.2. Processing Settings	57
2.9.3. Add Taxonomy	57
2.9.4. Review Your Configuration	58
3. Security (Users)	59
3.1. Secure Netwrix Data Classification	59
3.2. User Management	61
3.3. Password Manager	64
3.4. Web Service Security	65
4. Content Sources	66
4.1. Add a Content Source	66
4.1.1. Database	67
4.1.2. Exchange Mailbox	68
4.1.3. Exchange Server	69
4.1.4. File System	70
4.1.4.1. Add Folder source	70
4.1.4.2. Add Files source	72
4.1.5. Add Google Drive Source	73
4.1.6. Outlook Mail Archive	75
4.1.7. SharePoint	75
4.1.8. SharePoint Online	76
4.2. Narrow Data Collection Scope	77
4.3. Use Tagging (optional)	77
4.4. Manage Sources and Control Data Processing	78

4.4.1. Modify Source Settings	79
4.4.2. Set up granular processing and tagging for Database	80
4.4.3. Set up exclusions and tagging for Exchange	85
4.4.4. Set up filters and tagging for File System	85
4.4.4.1. Configure Inclusions	85
4.4.4.2. Configure Exclusions	86
4.4.4.3. Configure Tagging	87
4.4.5. Set up exclusions and tagging for Google Drive	89
4.4.6. Set up processing options for SharePoint	90
4.4.7. Set up processing options for SharePoint Online Tenancy	95
4.5. View Results	96
4.5.1. Data Processing Statistics	96
4.5.2. Content Crawling and Classification Results	96
5. Taxonomies	99
5.1. What are Taxonomies?	99
5.2. Built-in Taxonomies Overview	99
5.2.1. Core Taxonomies	100
5.2.2. Derived Taxonomies	102
5.3. Taxonomy Settings	104
5.3.1. Taxonomy Settings Levels	105
5.3.2. Labels	107
5.3.2.1. SharePoint Labels	108
5.3.2.2. O365 Labels	108
5.3.2.3. Help	108
5.4. Add a Taxonomy	108
5.5. Manage Taxonomies	109
5.5.1. Managing Term Sets	114
5.5.2. Multi-User Environments	114
5.6. Search and Filter Taxonomies	115
5.7. Classification Rules (Clues)	118

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

6.2.2.1. Specifying Rule Conditions	157
6.2.2.2. Specifying Rule Actions	159
6.2.2.3. Other Rule Settings	160
6.2.2.4. Specifying Workflow Conditions	160
6.2.3. Edit Workflow settings	162
6.2.4. Delete Workflow	163
6.3. Workflow Actions	164
6.3.1. Available Actions	164
6.3.1.1. Email Alert	165
6.3.1.2. Migrate Document	168
6.3.1.3. Apply Additional Classification	173
6.3.1.4. Advanced Actions for Exchange	174
6.3.1.5. Advanced Actions for File System	176
6.3.1.6. Advanced Actions for SharePoint	177
6.3.2. Plugins for Additional Actions	179
6.4. Workflow Operations Log	179
6.5. Workflow Plugins	180
7. Administrative Tasks	181
7.1. Index Maintenance	181
7.1.1. Step 1: Maintenance Operation	182
7.1.2. Step 2: Maintenance Options	182
7.1.3. Step 3: Summary	183
7.1.4. Step 4: Process	183
7.2. Configuration Options	183
7.2.1. Core Configuration	185
7.2.2. Licensing	185
7.2.3. Metadata Configuration	186
7.2.4. Email Configuration	188
7.2.5. Text Handling	190
7.2.6. Redaction	194

7.2.7. Additional Configuration Settings	195
7.2.8. Configuration Backup	198
7.3. Review Dashboards	200
7.3.1. System Health	201
7.3.2. Netwrix Data Classification Service Viewer	201
8. Reporting Capabilities	202
8.1. Content Distribution	203
8.2. Review Built-in Reports	203
8.3. Types of Reports	206

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:

Data Source	Supported Versions
File System	<ul style="list-style-type: none">• CIFS/SMB (Preferred)• NFS
SharePoint, SharePoint Online, OneDrive for Business	<ul style="list-style-type: none">• 2010 and above
Database	<ul style="list-style-type: none">• Microsoft SQL Server 2008 and above• Oracle 10g and above
Exchange	<ul style="list-style-type: none">• Exchange Server 2010 and above• Exchange Online
Google Drive	<ul style="list-style-type: none">• N/A
Outlook Mail Archive	<ul style="list-style-type: none">• Outlook 2010 and above

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](#).

2.2.2.0.2. 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:

Number of objects to classify	Environment	Comment
Up to 500, 000	Proof-of-concept and small-size environment	
Up to 8, 000, 000	Mid-size environment	
Up to 32, 000, 000	Large-size environment	
More than 32, 000, 000	Extra-large environment	System architect's assistance is required for deployment planning requires

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:

Number of objects to classify	Environment	Comment
Up to 500, 000	Proof-of-concept and small-size environment	
Up to 8, 000, 000	Mid-size environment	

Number of objects to classify	Environment	Comment
Up to 32, 000, 000	Large-size environment	
More than 32, 000, 000	Extra-large environment	System architect's assistance is required for deployment planning in such environments.

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

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

Specification	Settings	Comments
	<i>Hard disk:</i> SSD storage; thin provisioning enabled	
Guest OS	Windows Server 2012 R2 (64-bit)	
Software	Netwrix Data Classification server with <i>Distributed Query Server Mode</i> configuration.	Used 4 server instances with <i>Distributed Query Server Mode</i> configuration. See "Distributed Query Server Mode" for details.

Database Server

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

Data Source (File Server)

Specification	Settings	Comments
Platform	VMware ESXi 6.7	
Hardware:	<i>CPU:</i> Intel Xeon E5-2620 v4 , 2.10 GHz	Using faster processor increases data processing performance.

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

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.

	No Index	Keyword	Compound Term
Processing time	1 day 07 h 22 m	1 day 10 h 03 m	1 day 12 h 35 m

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

* — 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

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

Specification	Settings	Comments
Software	Netwrix Data Classification server with <i>Distributed Query Server Mode</i> configuration.	Used 4 server instances with <i>Distributed Query Server Mode</i> configuration. See "Distributed Query Server Mode" for details.

Database Server

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

Data Source (File Server)

Specification	Settings	Comments
Platform	VMware ESXi 6.7 Hardware: <i>CPU:</i> Intel Xeon E5-2620 v4 , 2.10 GHz <i>Logical Processors:</i> 32 (2 sockets, 8 cores per socket) <i>Memory:</i> 128 GB	Using faster processor increases data processing performance.

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

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.

	Keyword
Processing time	62 days 19 hrs 47 min
Files processed per minute (average)	365
Files with workflow condition triggered (i.e. at least 1 taxonomy applied)	4752724
NDC SQL database size (MDF)*	190GB
Index size	4 TB

* — 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.

Hardware Component	Minimum Requirements	Recommended
Processor	Any modern. Consider that greater CPU frequency and number of cores improve overall performance of Netwrix Data Classification Server.	Any multi-core
RAM	8 GB	16 GB

2.3.1.2. SQL Server

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

Hardware Component	Minimum requirements	Large environment (up to 8 m objects for File Servers and up to 2 m objects for SharePoint)	XLarge environment (up to 32 m objects and up to 8 m objects for SharePoint)
Processor	Any multi-core	8 cores	8 cores
RAM	16 GB	64 GB	128 GB
Disk space	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. See also Deployment Planning .		

NOTE: If you are going to process more than 10,000,000 objects per day, remember to perform database maintenance procedures. See [Recommendations on SQL Database Maintenance](#).

2.3.1.3. Network Access

Specification	Requirement
Network access	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.

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 the 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

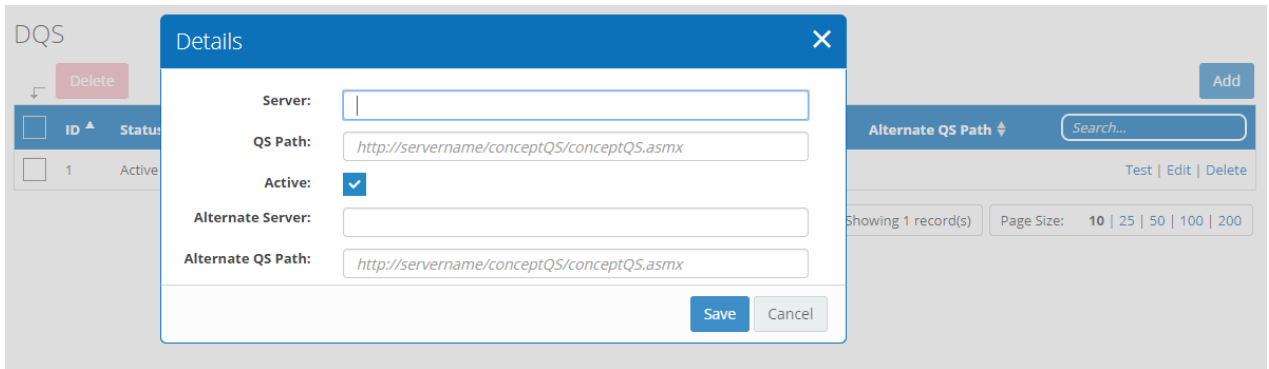
1. Install and configure the first Netwrix Data Classification Server as described in the [Install Netwrix Data Classification](#) section.
2. Open administrative web console.
3. Navigate to **Config** → **Utilities** → **DQS**.
4. Select **Enable DQS**.

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.

ID	Status	Load	Server	QS Path
1	Active	0%	FM-DDC-2	http://FM-DDC-2/conceptQS/conceptQS.aspx

Complete the following fields:



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.

Application Log
Backup/Restore
Cleaner
DQS
Search Log
Show User
Stem Test
Taxonomy Converter

DQS

Information

The Distributed Query Server (DQS) is a component of conceptClassifier that allows an index to be distributed across multiple servers.

A distributed index means that there are multiple servers running the Collector, Indexer, Classifier, and QueryServer applications - each with its own set of ".cse" files. All servers share a single SQL database.

Please note:

- Each server can only run one set of Windows Services
- The server name should be specified in the NETBIOS format (case insensitive)
- The QS Path specified should be a direct connection to the server in question (I.E. not a load balanced address for the cluster)
- The server should be set to "Active" to be considered part of the cluster
- All servers that you wish to run the Windows Services on should be specified within the DQS list

Delete

Add

<input type="checkbox"/>	ID	Status	Load	Server	QS Path	Alternate Server	Alternate QS Path	
<input type="checkbox"/>	1	Active	0%	FM-DDC-2	http://FM-DDC-2/conceptQS/conceptQS.aspx			Test Edit Delete
<input type="checkbox"/>	2	Active	0%	fm-ddc-4	http://fm-ddc-4/conceptQS/conceptQS.aspx			Test Edit Delete
<input type="checkbox"/>	3	Active	0%	fm-ddc-5	http://fm-ddc-5/conceptQS/conceptQS.aspx			Test Edit Delete
<input type="checkbox"/>	4	Active	0%	fm-ddc-3	http://fm-ddc-3/conceptQS/conceptQS.aspx			Test Edit Delete

Copy | CSV | XLSX
Showing 4 record(s)
Page Size: 10 | 25 | 50 | 100 | 200

- 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:

Component	Requirements
Operating system	Windows 2012 R2 and above Server Operating System Software.
Windows Features	<div>Web Server Role (IIS)</div> <hr/> <div>Common HTTP Features</div> <ul style="list-style-type: none"> • Default Document • HTTP Errors • Static Content • HTTP Redirection <hr/> <div>Security</div> <ul style="list-style-type: none"> • Windows Authentication • Anonymous Authentication <div> NOTE: The Anonymous Authentication element is included in the default installation of IIS 7. Make sure you use IIS 7 and above. </div> <hr/> <div>Application</div> <ul style="list-style-type: none"> • ISAPI Extensions <div>Development</div> <ul style="list-style-type: none"> • ISAPI Filters <hr/> <div>Other features</div> <hr/> <div>.NET Framework</div> <ul style="list-style-type: none"> • .NET Framework 4.7.2 <div>Features</div> <ul style="list-style-type: none"> • ASP.NET <hr/> <div>WCF Services</div> <ul style="list-style-type: none"> • HTTP Activation
SQL Server	<ul style="list-style-type: none"> • SQL Server 2008 R2 Standard Edition (or later). • SQL Server 2016 SP2 recommended (for better performance). <p>NOTE: For large environments, SQL Server Enterprise edition may be needed; see needed. See Deployment Planning.</p>

Component	Requirements
Microsoft IFilterers	<ul style="list-style-type: none">• Microsoft Office 2010 Filter Packs and above, 64-x edition.
Visual Studio	<ul style="list-style-type: none">• Visual C++ Redistributable Packages for Visual Studio 2015 and above.

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 <i>.CSE</i> extension.

2.3.3. Accounts and Required Permissions

Netwrix Data Classification uses the following accounts:

Account	Description
Service Account	<p>This account is specified during the product setup.</p> <p>Windows domain account that you plan to use as a service account will need the following:</p> <ul style="list-style-type: none"> • Local Administrator rights on the server where Netwrix Data Classification will be installed. • Permissions to run the Windows Services and IIS Application pool. <p>After installation, this account will be automatically granted the Logon as a service privilege on the Netwrix Data Classification server.</p> <ul style="list-style-type: none"> • SQL Server DBO permissions to the NDC SQL database (if using Windows Authentication to access SQL Server). <p>NOTE: Optionally, you can use local account instead of domain account.</p>
Crawl content	<p>Ensure the availability of accounts with sufficient permissions to access your content sources:</p> <ul style="list-style-type: none"> • SharePoint, SharePoint Online site collection— Site Collection Administrator role. • Exchange mailboxes: <ol style="list-style-type: none"> 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. <p>See Configure Microsoft Exchange for Crawling and Classification for detailed information on configuring these permissions.</p> <ul style="list-style-type: none"> • Outlook Mail Archive (PST file)— Read permission. • File System (SMB, NFS) — Read permission for the folders and files you need to crawl.

Account	Description
	<ul style="list-style-type: none">• 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. <p>See Configure G Suite for Crawling for detailed information.</p> <ul style="list-style-type: none">• Database— Read permission for the database schema and data.
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:

Option	Description
File Growth	<ul style="list-style-type: none">• Recommended—128 MB.• Large environment— 512 MB.
Maximum File Size	Select Unlimited .

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:

Option	Description
Server Name	Provide the name of the SQL Server instance that hosts your NDC SQL database. For example, "WORKSTATIONSQ\SQLSERVER".
Authentication Method	Select Windows or SQL Server authentication method.
Username	Specify the account name.
Password	Provide your password.
Database Name	Enter the name of the SQL Server database. Netwrix recommends using NDC_database name.

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](#).
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

Name	Group Name	Status	Location	Username	Actions
CCPA	aff0c9a7-4115-4878-9f0a-d77e3b63baf1	Online	SQL		Update Compare Edit Delete
File Size	c35a73d9-4ba1-432a-97a0-5993ae18b760	Online	SQL		Update Compare Edit Delete
File Type	ff0278af-0ff7-4c37-b569-3d3a4dc8919c	Online	SQL		Update Compare Edit Delete
Financial Records	2844625c-5c81-459a-805a-ea880a70a67	Online	SQL		Update Compare Edit Delete
GDPR Restricted	baf14908-8860-4856-9278-cc5f93a623d	Online	SQL		Update Compare Edit Delete
GDPR	3c7a922a-8a46-432f-a166-82ac1aff9a2c	Online	SQL		Update Compare Edit Delete
GLBA	0a8057d2-197b-460b-8a4d-278a6f312a08	Online	SQL		Update Compare Edit Delete
HIPAA	0a2b780a-4498-4bca-bc2c-39229a379aaf	Online	SQL		Update Compare Edit Delete
Language	d72878a4-3f80-4966-8ccc-b38026c0d043	Online	SQL		Update Compare Edit Delete
PCI DSS	0f0ab07a-30a6-4151-9353-eecc0b0c0188	Online	SQL		Update Compare Edit Delete

- After the upgrade, indexing mode will be set to **Compound Term** mode. Refer to the following Netrix 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 Netrix 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 the 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

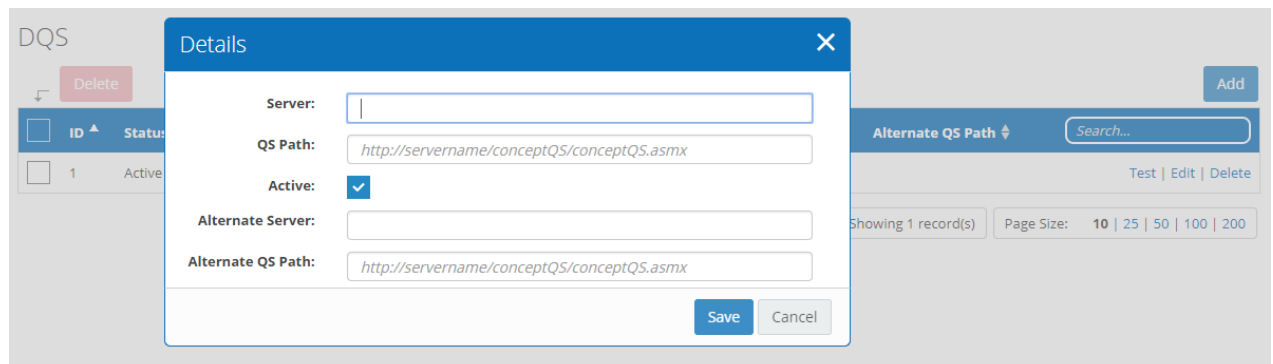
1. Install and configure the first Netwrix Data Classification Server as described in the [Install Netwrix Data Classification](#) section.
2. Open administrative web console.
3. Navigate to **Config** → **Utilities** → **DQS**.
4. Select **Enable DQS**.

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:



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.

- Click **Save** to close the dialog.
 - 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](#)
 - On each server, follow the installation steps as described in the [Install Netwrix Data Classification](#) section until **SQL Database** configuration.
 - 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.
- Complete the installation.
 - Repeat steps 2 - 6 for every NDC Server, then review the list of servers to make sure the new server was included.

Application Log
Backup/Restore
Cleaner
DQS
Search Log
Show User
Stem Test
Taxonomy Converter

DQS

Delete
Add

ID	Status	Load	Server	QS Path	Alternate Server	Alternate QS Path	
1	Active	0%	FM-DDC-2	http://FM-DDC-2/conceptQS/conceptQS.aspx			Test Edit Delete
2	Active	0%	fm-ddc-4	http://fm-ddc-4/conceptQS/conceptQS.aspx			Test Edit Delete
3	Active	0%	fm-ddc-5	http://fm-ddc-5/conceptQS/conceptQS.aspx			Test Edit Delete
4	Active	0%	fm-ddc-3	http://fm-ddc-3/conceptQS/conceptQS.aspx			Test Edit Delete

Copy | CSV | XLSX
Showing 4 record(s)
Page Size: 10 | 25 | 50 | 100 | 200

Information

The Distributed Query Server (DQS) is a component of conceptClassifier that allows an index to be distributed across multiple servers.

A distributed index means that there are multiple servers running the Collector, Indexer, Classifier, and QueryServer applications - each with its own set of ".cse" files. All servers share a single SQL database.

Please note:

- Each server can only run one set of Windows Services
- The server name should be specified in the NETBIOS format (case insensitive)
- The QS Path specified should be a direct connection to the server in question (i.e. not a load balanced address for the cluster)
- The server should be set to "Active" to be considered part of the cluster
- All servers that you wish to run the Windows Services on should be specified within the DQS list

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:

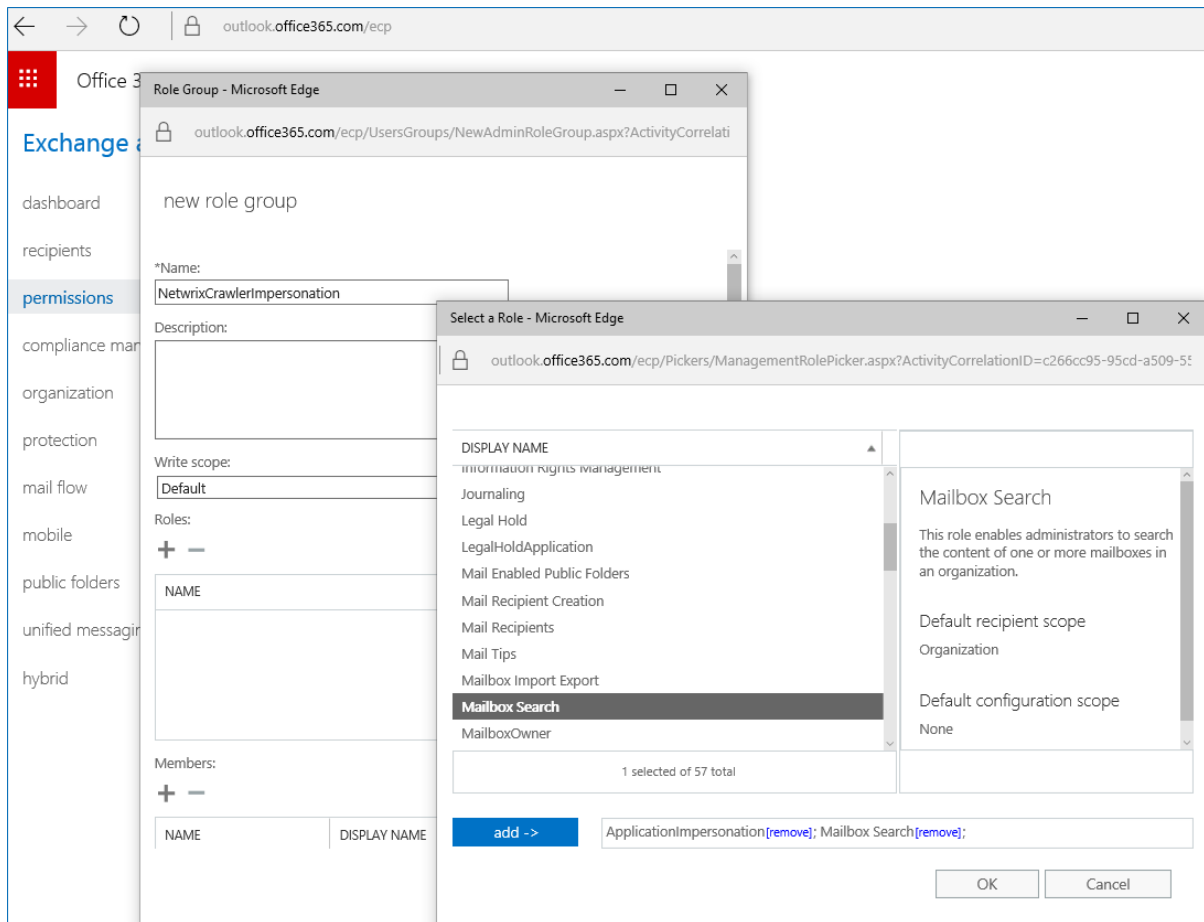
- **ApplicationImpersonation**—Allows the crawling account to impersonate each of the mailboxes / users configured for collection
- **Mailbox Search**—Allows the crawling account to enumerate mailboxes (automatic discovery of mailboxes)

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](#)
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):

```
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:

To...	Do...
Create a new project	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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.

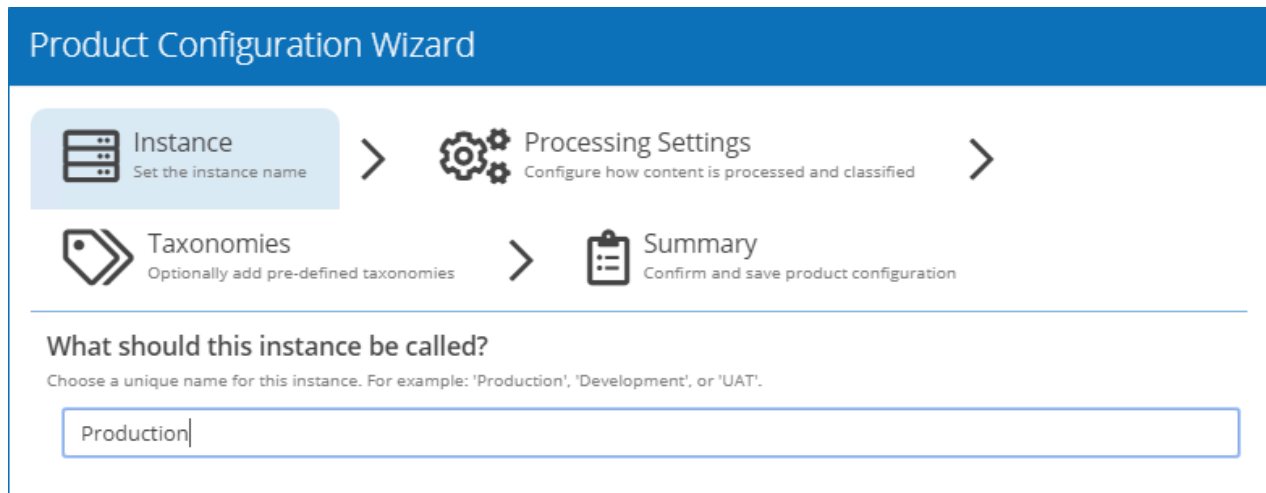
To...	Do...
Create a service account key	<ol style="list-style-type: none"> 1. On the Service accounts page, select the account you want to create a key for. 2. Click  icon under Actions and select Create key. 3. In the Create private key for <Service account name> dialog, select JSON format, and download the file to a known location as it will be required later. <p>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.</p>
Delegate domain-wide authority to the service account	<ol style="list-style-type: none"> 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 "<i>Domain wide delegation</i>" 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	<ol style="list-style-type: none"> 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 "<i>https://www.googleapis.com/auth/drive</i>" 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"*.



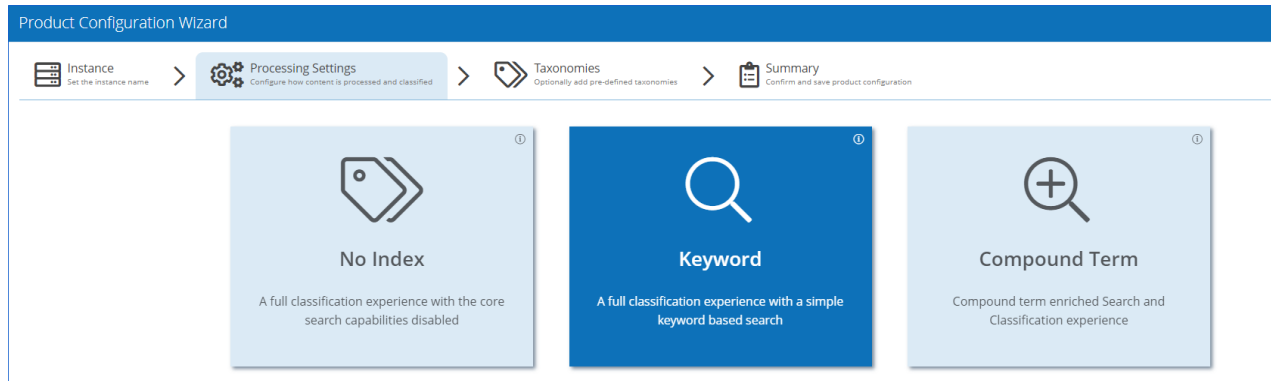
The screenshot shows the 'Product Configuration Wizard' interface. At the top, there's a blue header with the title 'Product Configuration Wizard'. Below the header, there's a horizontal navigation bar with four steps: 'Instance' (Set the instance name), 'Processing Settings' (Configure how content is processed and classified), 'Taxonomies' (Optionally add pre-defined taxonomies), and 'Summary' (Confirm and save product configuration). The 'Instance' step is currently selected and highlighted. Below the navigation bar, there's a section titled 'What should this instance be called?' with a subtext: 'Choose a unique name for this instance. For example: 'Production', 'Development', or 'UAT''. A text input field contains the word '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.

The screenshot shows the 'Processing Settings' step of the 'Product Configuration Wizard'. It contains the following sections and options:

- Text Extraction**
 - Should OCR be used on image files?**

OCR is used to extract text from images. This is useful if the content being collected contains a large number of scanned documents (for example), image file extensions will be automatically added to the list of "Files Included" if this setting is enabled.

☒ Yes ☐ No

Information
OCR requires the Visual C++ Redistributable for Visual Studio 2015, which is available from the following [link](#).
 - Should images embedded in documents be processed?**

Images inside office documents (e.g., DOC and XLS files) or PDF files can be processed using OCR. Any text extracted will be appended to the document text. Note that this option can dramatically affect the processing speed of content.

☐ Yes ☒ No
 - Should the collection process optimise text storage by re-using text offsets?**

This reduces the storage requirements for the local database (stored text) by sharing and reusing the stored text when matches are identified. However, this does result in a small increase in sql database demands.

☐ Yes ☒ No
- Classification Configuration**
 - Should default clues be automatically created?**

When enabled a clue will automatically be created when a taxonomy is registered from SharePoint or a term is created. The new clue will either be a standard clue matching the term name or a metadata clue depending on the configuration specified at the taxonomy level settings.

☐ Yes ☒ No
 - Should boosted phrasematch scoring be enabled?**

When switched on, the score of any phrasematch clues will be boosted if the phrase appears multiple times in the document.


☒ Yes ☐ No

Proceed with adding taxonomies.


2.9.3. Add Taxonomy

On this step, you are prompted to load predefined taxonomies.


Product Configuration Wizard

 Instance
Set the instance name


>

 Processing Settings
Configure how content is processed and classified

>

 Taxonomies
Optionally add pre-defined taxonomies

>

 Summary
Confirm and save product configuration

Which preloaded taxonomies would you like to load?

These taxonomies come pre-populated with terms/clues and can be deleted and reloaded as required

✕ GDPR

✕ HIPAA

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.

Users

Users > Add User

User Details

Username

Superuser ☒ (mandatory for first user)

Allow Rest API Access ☐

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 back-end 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/"`.

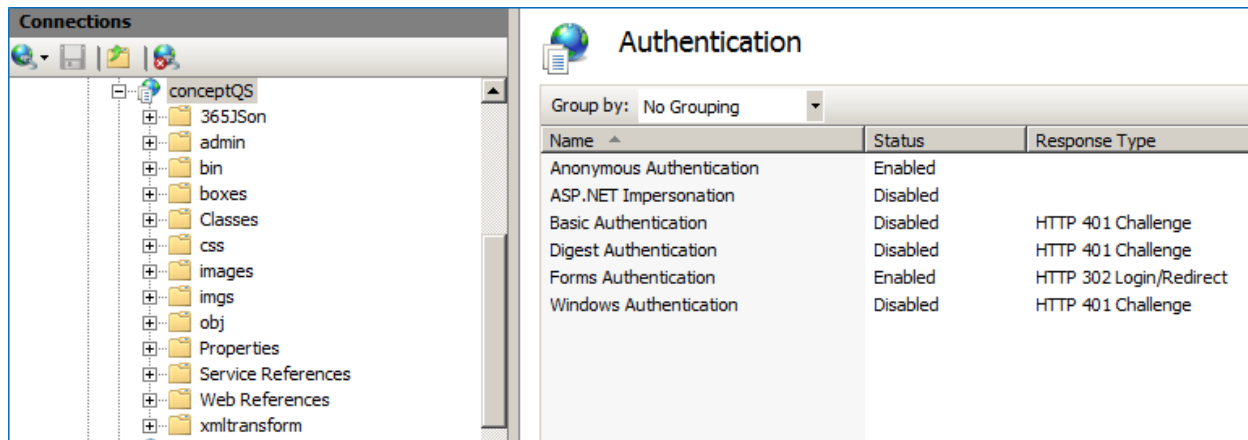
3.2. User Management

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.

Users

Users > Add User

User Details

Username:

Superuser: ☒ (mandatory for first user)

Allow Rest API Access: ☐

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.

The screenshot shows the 'Users' management interface. At the top, the breadcrumb is 'Users > FMWSAPERF\Test'. Below it, there are tabs: 'User Details', 'Sources' (selected), 'Taxonomies', 'Workflows', 'Config', 'Users', and 'Reports'. Under the 'Sources' tab, there is a checkbox 'Access Sources' which is checked. To the right of this is a 'Save Permissions' button. Below these are four panels with blue headers:

- Allowed Source Groups:** Contains a checkbox 'All Source Groups' which is checked.
- Source Management:** Contains four checkboxes: 'Add', 'Edit', 'Edit Advanced Config', and 'Delete'.
- Page Management:** Contains one checkbox: 'Delete'.
- Actions:** Contains four checkboxes: 'Pause / Resume', 'Re-Collect', 'Re-Index', and 'Re-Classify'.

Taxonomy Permissions Summary:

View Taxonomy Permissions (2)
Save Permissions

Users

[-] Add
[+] Edit

Actions

☐ Rollback

Permissions
✕

Level ▲	Permissions ◆
Taxonomy: IPSV v2	Term (Add, Edit)
Taxonomy: Sensitive Information	Term (Add)

📄 Copy | CSV | XLSX

Showing 2 record(s)

Page Size: 10 | 25 | 50 | 100 | 200

Cancel

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.

Password Manager

Username ^ Domain ^ Reference ^ Auth Type ^ Auth Server ^					Add	
cs-admin	conceptdemo		AD		Manage Passwords Edit	
					Showing 1 record(s) Page Size: 10 25 50 100 200	

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.

Web Service Security

Method Group Security			Manage Specific Methods	
Group Title ^	Security Status ^	Allowed Users ^	Search...	
conceptClassifier Wizard/Templating Configuration Services	✓ Allow All		View Methods Edit	
ConceptSearching Internal Webservice Functions	✓ Allow All		View Methods Edit	
ConceptSearching System/Instance Information	✓ Allow All		View Methods Edit	
CS Index/SQL Read Actions	✓ Allow All		View Methods Edit	
CS Index/SQL Write Actions	✓ Allow All		View Methods Edit	
Query Server Admin Services	✓ Allow All		View Methods Edit	
Taxonomy Read	✓ Allow All		View Methods Edit	
Taxonomy Write	✓ Allow All		View Methods Edit	
Update instance settings functions	✓ Allow All		View Methods Edit	
Utility Functions	✓ Allow All		View Methods Edit	
			Showing 11 record(s) Page Size: 10 25 50 100 200	

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:

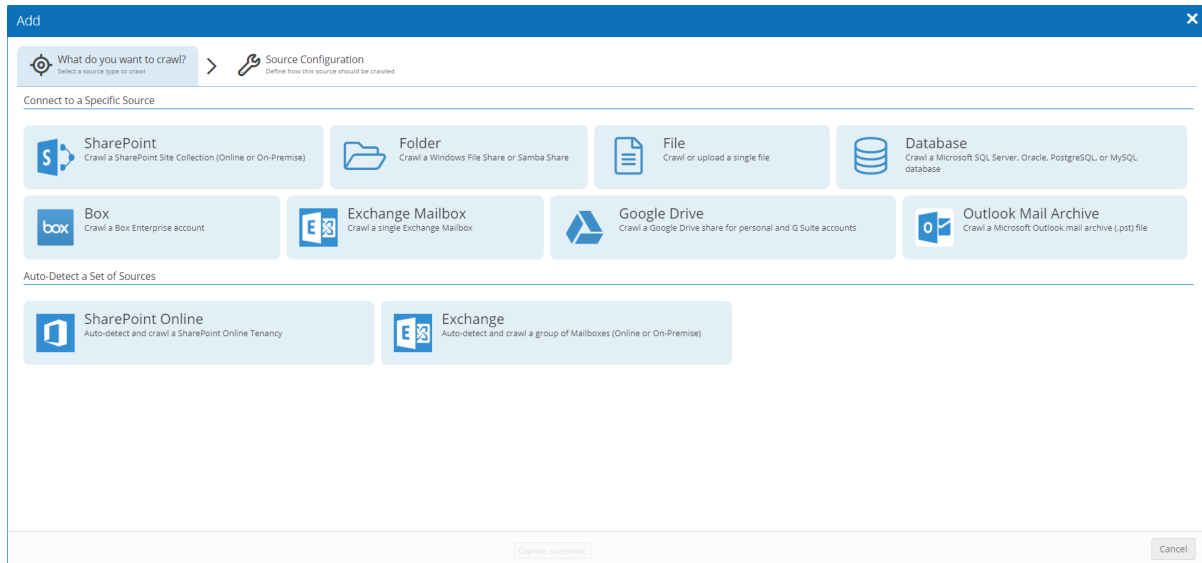
- [Add a Content Source](#)
- [Modify Source Settings](#)
- [Manage Sources and Control Data Processing](#)
- [Source Groups](#)

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:

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

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:

Option	Description
Email Address / Password	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 .

Option	Description
	By default crawling will attempt to locate the correct Exchange URL by using the Exchange Auto Discover functionality. Where this is not available the API URL should be specified. This is typically in the format: <i>https://servername/EWS/Exchange.asmx</i> .
Mailbox	When using impersonation the Mailbox should be specified as the mailbox to be crawled, for example: Email Address set to administrator@cs.com, and Mailbox set to test@cs.com.
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.
Re-Index Period	Specifies how often the source should be checked for changes. The number specifies the period in days.
Build Search index	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.
Document Type	Used to specify a value which can be used to restrict queries when utilizing the Netwrix Data Classification search index.
Pause source on creation	Select if you want to make other configuration changes before collection of the source occurs.

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.

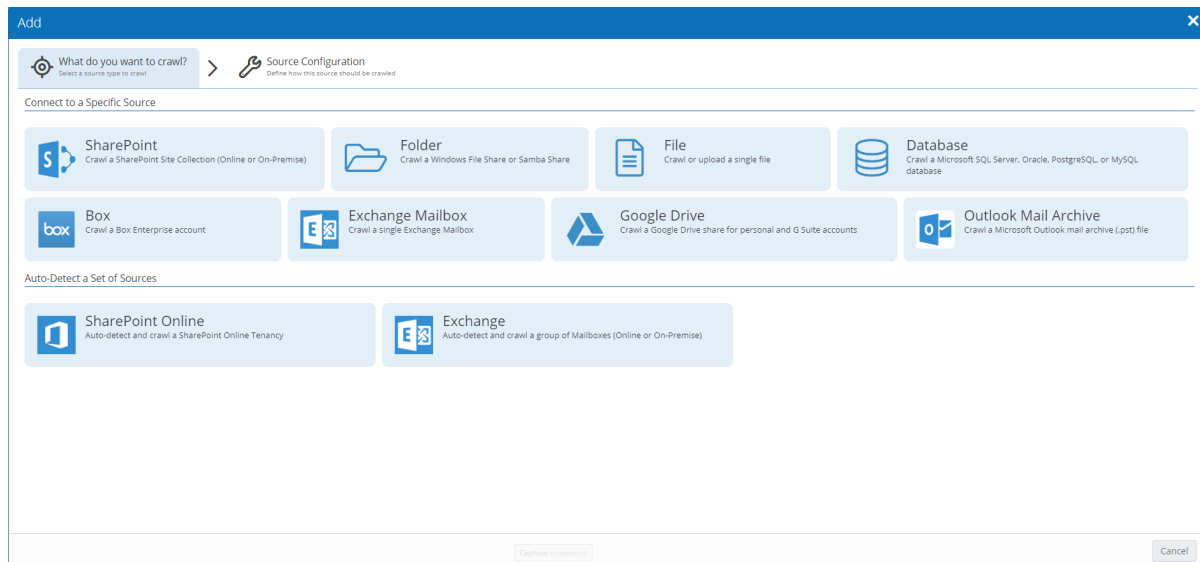
Complete the following fields:

Option	Description
Email Address / Password	<p>The Email Address / Password combination must be an administrator account that has been assigned the right of Impersonation as well as the Discovery Management role.</p> <p>By default crawling will attempt to locate the correct Exchange URL by</p>

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: <i>https://servername/EWS/Exchange.asmx</i> .
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:

Option	Description
Basic settings	
Folder	Enter the UNC path of the root folder where collection is to start.
Depth Limit	Specify how many levels the indexing should process. Possible options: <ul style="list-style-type: none"> • Exclude Subfolders • All Subfolders (default setting) • Limit Subfolders - if selected, specify the required subfolders depth (from 2 to 99)
Write classifications	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. See also Set up filters and tagging for File System .
Source Group	Default value recommended.
Pause source on creation	Select if you want to make other configuration changes before collection of the source occurs.
Advanced settings	
Username	Specify the account used to process the folder.
Password	Provide a password for the account specified above.
Text Patterns	See Text Handling for more information.
Date Filter	Use this calendar control to instruct the program to only crawl the content that has been modified since the specified date. 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).

Option	Description
Anonymous Access Allowed	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.
Duplicate Detection Enabled	Select to exclude duplicates (i.e. documents that contain the same text content) from the index.
Re-Index Period	Specifies how often the source should be checked for changes. Netwrix recommends using default values. Default is 7 days .
Priority	Netwrix recommends using default values.
Document Type	Specify a value that will be used to restrict queries when utilising the search index.

When finished, click **Save**.

4.1.4.2. Add Files source

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

The screenshot shows the 'Source Configuration' window with the 'File' source selected. The 'File Source' section has two radio buttons: 'File' (unselected) and 'Browse' (selected). Below this is a 'File' input field with a 'Browse...' button. The 'Source Group' dropdown menu is set to 'Additional 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.

Option	Description
Basic settings	
File Source	Select how you wish to provide the file location: <ul style="list-style-type: none"> ◦ File - enter file path

Option	Description
	<ul style="list-style-type: none"> ◦ Browse - browse for the file you need
Source Group	Default value recommended.
Advanced settings	
Username	Specify the account used to process the file.
Password	Provide a password for the account specified above.
Anonymous Access Allowed	<p>Select this option to disable security filtering for the content source.</p> <p>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.</p>
Upload	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.
Text Patterns	See Text Handling for more information.
Max Collector Retries	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 3 retries.
Re-Index Period	Specifies how often the source should be checked for changes. Netwrix recommends using default values. Default is 7 days .
Priority	Netwrix recommends using default values.
Document Type	Specify a value that will be used to restrict queries when utilising the search index.

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.

Add

What do you want to crawl?
Select a source type to crawl

>

Source Configuration
Define how this source should be crawled

Drive Type: ⓘ ☐ Personal ☒ Business

User Email(s):

✕ administrator@corp.local

✕ security.team@corp.local

Crawl Shared Items: ⓘ ☒

JSON Import:

Drag and drop a JSON file containing the service account credentials here, or copy the JSON t

Project ID:

Click here to show/hide all configuration fields for Google Drive

Write Classifications: ☐ ⓘ **Note: Document audit information will be altered**

Source Group:

Pause source on creation: ☒

Complete the following fields:

Option	Description
Basic settings	
Drive Type	Select <i>Business</i> .
User Email(s)	When adding a G-Suite source, enter the email address of the user's drive that you wish to crawl (via impersonation).
Crawl Shared Items	Select to crawl all files shared with the specified user in addition to any team drives shared with the user.
Crawl Shared Items	Select to enable crawling of any types of documents shared with the specified user.
JSON Import	Drag the JSON connection file you downloaded while creating Google service account in the form.
Project ID	Open the JSON connection file and copy file contents to Project ID field.
Write Classifications	Leave this checkbox empty.

Option	Description
Source Group	Netwrix recommends creating a dedicated source group for Google Drive.
Pause source on creation	Select if you want to make other configuration changes before collection of the source occurs.

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**.

The screenshot displays the 'Source Configuration' interface for an Outlook Mail Archive. It features two main sections: 'What do you want to crawl?' and 'Source Configuration'. The 'What do you want to crawl?' section includes a 'File:' input field with a '+' button. The 'Source Configuration' section includes a 'Source Group:' dropdown menu currently set to 'None', with a '+' and refresh icon to its right, and a 'Pause source on creation:' checkbox.

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**.

What do you want to crawl?
Select a source type to crawl

>

Source Configuration
Define how this source should be crawled

URL:

Username:

Password:

Write classifications: ☐

Source Group:

Pause source on creation: ☐

Complete the following fields:

Option	Description
SharePoint URL	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.
Username	Enter username in the following formats: DOMAIN\USERNAME and USERNAME@DOMAIN.
Write Classifications to SharePoint	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.
Re-Index Period	Specifies how often the source should be checked for changes. The number specifies the period in days.
Document Type	Specify a value which can be used to restrict queries when utilizing the Netwrix Data Classification search index.

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 One Drive 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.

What do you want to crawl?
 Select a source type to crawl

Source Configuration
 Define how this source should be crawled

URL:

Username:

Password:

We recommend that the crawling account is a tenancy administrator. During crawling we will automatically enable the user as a site collection administrator on any detected site collection(s). This is to ensure that we have the required permissions for both crawling and writing classifications back to SharePoint.

Match Rules:

At least one match rule must be included, match rules are Regular expressions, such as:


```
s:
https://conceptsearching.sharepoint.com/sites/V/*
.*VPersonal/V.*
https://conceptsearching-my.sharepoint.com/V.*
```

+

No values

Classification Template:

Disabled (No classifications will be written to SharePoint)

▼

Detection Period:

0 day(s) (disabled)

0 hour(s)

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

- Enabling external workflow applications
- Applying IT policies to the classified objects

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

Netwrix Data Classification 5.5.1							Sources Taxonomies Workflows Config Users Reports Dashboard Help							
General							Box	CMIS	Content Server	Exchange Mailbox	File	Google Drive	SharePoint	Web
Sources														
<div> Delete Re-Collect Re-Index Re-Classify Pause Resume Add To Group Add </div>														
<input type="checkbox"/>	Page Url	Status	Collect Content	Build Index	Documents	Size	Search...							
<input type="checkbox"/>	\\wsaperf-t210\testshare	Processed	✓	✓	1061	684 KB	⚙️ 📊 ℹ️ 🔗							
<input type="checkbox"/>	gdrive://netwrixclassifier	Processed	✓	✓	1067	3.93 MB	⚙️ 📊 ℹ️ 🔗							
Copy CSV XLSX Showing 2 record(s) Page Size: 10 25 50 100 200 Find Page														

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:

- [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.

Advanced SQL Configuration

Sources > Server=., Database=conceptQS

Entity Configuration

Add Query

Name	Type	Enabled	
dbo.[SecurityAudit]	Table	✗	Edit
dbo.ACLGroupMembership	Table	✗	Edit
dbo.ACLs	Table	✗	Edit
dbo.ACLUserMembership	Table	✗	Edit
dbo.AnalyticsQueue	Table	✗	Edit
dbo.ApplicationLog	Table	✗	Edit
dbo.Attachments	Table	✗	Edit
dbo.AttachmentsExcluded	Table	✗	Edit
dbo.AutoClassificationChanges	Table	✗	Edit
dbo.Backups	Table	✗	Edit

1 2 3 4 5 ... 16
 [Copy](#) | [CSV](#) | [XLSX](#)
 Showing 157 record(s)
 Page Size: 10 | 25 | 50 | 100 | 200

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:

Add Query

Name:

Example Query

Examples:

None

Primary Key Query:

SELECT ID FROM Docs

Content Query:

SELECT * FROM Docs

Save

Cancel

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

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:

Query		Edit
Primary Key Query:	<div>SELECT ID FROM Orders WITH(NOLOCK)</div>	
Content Query:	<div>SELECT * FROM Orders</div>	

Table Configuration

The table configuration allows you to choose how each specific entity will be crawled:

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

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 (Incremental Crawls)	Filter	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.	

dbo.CachedDocuments

Details

Include for Crawling:

☐

Modified Filter (Incremental Crawls):

-

Re-Index Period:

1

Days

0

Hours

0

Mins

Column Mappings

Mark all table columns as content:

☐

Mark all table columns as metadata:

☐

Field Name	Type	Primary Key	Content	Metadata
CachedDateTime	datetime	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CachedDocumentId	int	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FileContent	varbinary	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FileName	nvarchar	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
InitialContentTypeId	varchar	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ItemId	int	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ListId	varchar	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
WebId	varchar	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Index Mappings

Index Field	Mapped Field
ContentFilename	-
ContentType	-
DocDate	-
DocSummary	-
DocType	-
DocumentId1	-
DocumentId2	-
DocumentId3	-
DocumentId4	-
DocumentId5	-

Save

Cancel

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

Option	Description
	same as the table being crawled).
Column Name	Specify the name of the column to be updated (text/varchar column).
Update Filter	<p>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.</p> <p>The specified values will result in a query in the following format:</p> <pre>UPDATE TABLENAME SET COLUMNNAME=@Classifications WHERE FILTERS</pre>
Format	Specify the delimiters/construction of the value to be written into the SQL database.

Write Configuration

✕

Enabled: ☒

Table name:

Docs

For example in SQL Server:

[TableName]

[Schema].[TableName]

Column Name:

[ColumnToUpdate]

Update Filters:

✕ Title=@PageTitle

+

Format:

Delimited (Label | ClassId;Label | ...)

Name/Id:

|

Class:

;

Prefix:

Suffix:

Save

Cancel

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.

- To modify an extension (for example, add a wildcard), click **Edit**. To add a new one, click **Add**.

4.4.4.2. Configure Exclusions

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

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

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

Agriculture [X]

Enabled: ☒

Field Name: [Help]

Format:

Name/Id:

Class:

Prefix:

Suffix:

Save **Cancel**

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.

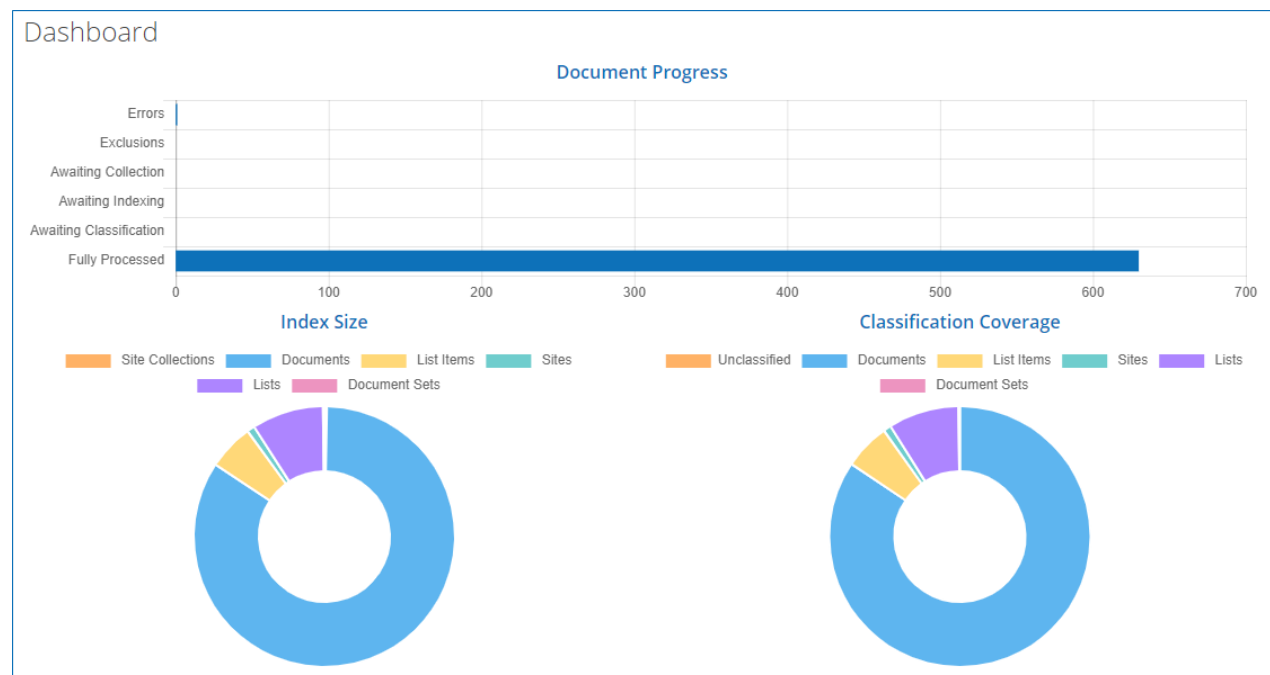
4.4.6. Set up processing options for SharePoint

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](#)
- [Advanced Configuration](#)
- [Entity Configuration Grid](#)
- [List Configuration](#)
- [Subsite Configuration](#)
- [Source Configuration \(Defaults\)](#)
- [SharePoint Excluded](#)
- [Templating](#)

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

- **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.

Advanced SharePoint Configuration

Sources > https://conceptsearching.sharepoint.com

Entity Configuration Source Defaults Configuration Viewer

Navigate: Root

Title	URL	Include	Has Config?	
Search Center	https://conceptsearching.sharepoint.com/Search Center	✓	✗	Exclude Edit
Financial Reports	https://conceptsearching.sharepoint.com/Financial Reports	✓	✗	Exclude Edit
Medical	https://conceptsearching.sharepoint.com/Medical	✓	✗	Exclude Edit
PII Cloud	https://conceptsearching.sharepoint.com/PII Cloud	✓	✗	Exclude Edit
Site Assets	https://conceptsearching.sharepoint.com/SiteAssets	✓	✗	Exclude Edit
Site Collection Documents	https://conceptsearching.sharepoint.com/SiteCollectionDocuments	✓	✗	Exclude Edit
Site Collection Images	https://conceptsearching.sharepoint.com/SiteCollectionImages	✓	✗	Exclude Edit
Site Pages	https://conceptsearching.sharepoint.com/SitePages	✓	✗	Exclude Edit
Style Library	https://conceptsearching.sharepoint.com/Style Library	✓	✗	Exclude Edit

Copy | CSV | XLSX Showing 9 record(s) Page Size: 10 | 25 | 50 | 100 | 200

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.

General

Document Date: Modified (Modified)

Content Fields: Please Select

No values

Special Field Mappings

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 subweb level settings and finally the source level settings.

Content Type:

Title: Title (Title)

Language:

Summary:

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 subweb level settings and finally the source level settings.

Content Type:	<input type="text"/>
Title:	<input type="text"/>
Language:	<input type="text"/>
Summary:	<input type="text"/>
Document Type:	<input type="text"/>
DocumentId3:	<input type="text"/>
DocumentId4:	<input type="text"/>
DocumentId5:	<input type="text"/>
DocumentId6:	<input type="text"/>
DocumentId7:	<input type="text"/>
DocumentId8:	<input type="text"/>
MetaAppend:	<input type="text"/>

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).


Advanced SharePoint Configuration

Sources > <https://conceptsearching.sharepoint.com>

Entity Configuration | Source Defaults | Configuration Viewer


General

Re-Index Period: Days: Hours: Mins:

Text Patterns: 

Write classifications? ☐

Date Field Mappings

Document Date: 

Backup Document Date:

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).

Content Field Mappings

Defining content field mappings provides a simple way to assign additional SharePoint fields to the text indexed as part of the document text. Mappings of each type will be taken first from the list level settings, before reverting to the subweb level settings and finally the source level settings.

Document Library Content Fields: Please Select
 X Browser Title (SeoBrowserTitle)

Pages Libraries Content Fields: Please Select
 X Account (Name) X Active (ChannelsActive)

Lists Content Fields: Please Select
 No values

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:

SharePoint Excluded

Exclusions are case-insensitive and can be either an exact match or a partial match by starting and/or ending the exclusion filter with a '*'.

Delete Test Add

<input type="checkbox"/> Value ^	Search...
<input type="checkbox"/> https://excludedsite.sharepoint.com/*	Edit Delete

Copy | CSV | XLSX Showing 1 record(s) Page Size: 10 | 25 | 50 | 100 | 200

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. 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.

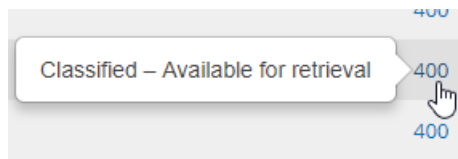
Sources

Sources > <https://conceptsearching.sharepoint.com> > Concept Searching Team Site > Shared Documents

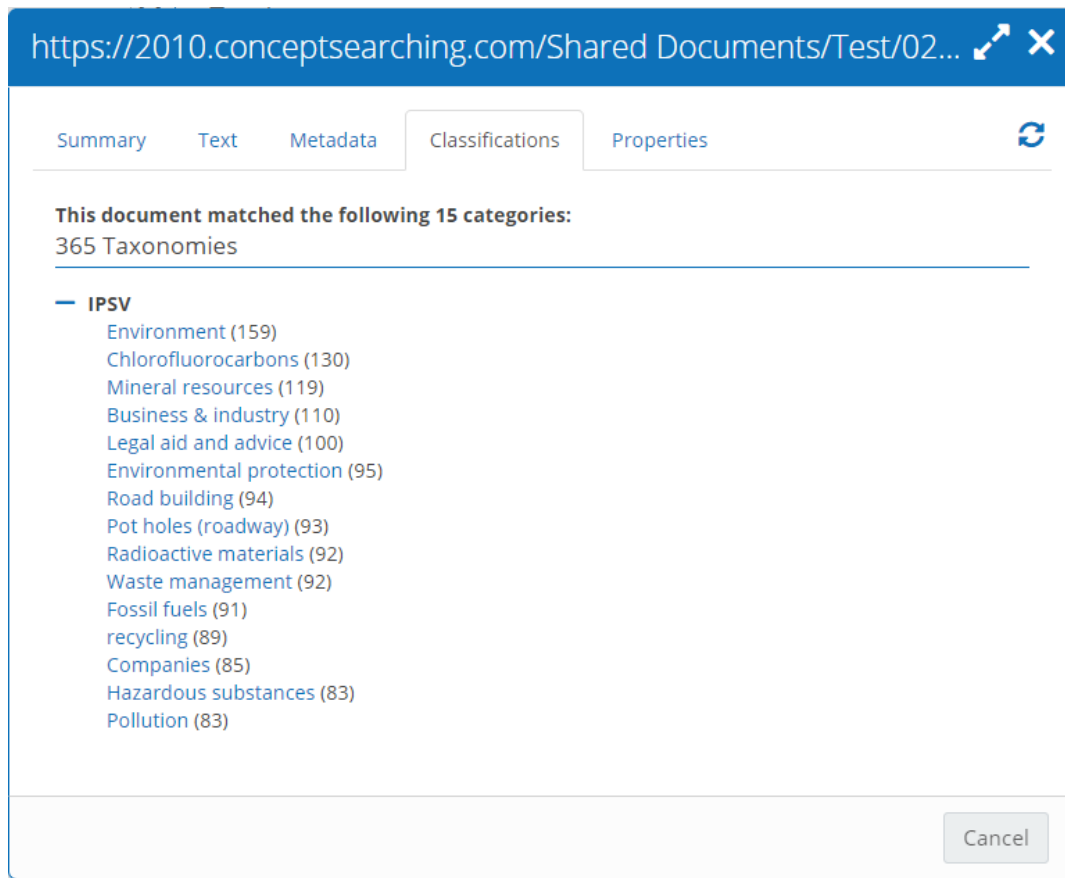
[Delete](#)
[Re-Index](#)
[Re-Classify](#)

<input type="checkbox"/>	Page Url	Status		
<input type="checkbox"/>	https://conceptsearching.sharepoint.com/Shared Documents/newsletter 0607.pdf	400		
<input type="checkbox"/>	https://conceptsearching.sharepoint.com/Shared Documents/newsletter 0608.pdf	400		
<input type="checkbox"/>	https://conceptsearching.sharepoint.com/Shared Documents/newsletter 0610.pdf	400		
<input type="checkbox"/>	https://conceptsearching.sharepoint.com/Shared Documents/newsletter 0611.pdf	400		
<input type="checkbox"/>	https://conceptsearching.sharepoint.com/Shared Documents/newsletter 0612.pdf	400		
<input type="checkbox"/>	https://conceptsearching.sharepoint.com/Shared Documents/newsletter 0613.pdf	400		
<input type="checkbox"/>	https://conceptsearching.sharepoint.com/Shared Documents/newsletter 0614.pdf	400		
<input type="checkbox"/>	https://conceptsearching.sharepoint.com/Shared Documents/Oil Spill.docx	400		
<input type="checkbox"/>	https://conceptsearching.sharepoint.com/Shared Documents/Pages from US5392735_Patent.pdf	400		

- 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.

The screenshot shows the 'Environment' taxonomy page. On the left, a sidebar lists various taxonomies under 'IPSV', with 'Environment (3)' selected. The main area displays a table of clues for the 'Environment' taxonomy. The table has columns for Type, Clue, #, and Score. The clues listed are:

Type	Clue	#	Score
Standard	Natural resources Languages	40	-
Standard	environmental protection Languages	30	-
Standard	climate Languages	20	-
Standard	pollution Languages	20	-
Standard	conservation Languages	10	-
Standard	protection Languages	10	-
Standard	Wildlife Languages	10	-
Standard	Environment Languages	10	-

At the bottom of the table, it says 'Showing 8 record(s)' and 'Page Size: 10 | 25 | 50 | 100 | 200'.

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

- 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, disease 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
- Luxembourg
- 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

- Latvia
- Lithuania
- Luxembourg
- 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](#)

- [Multi-User Environments](#)

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.

Taxonomy Settings

Taxonomy Name:

IPSV v2

Taxonomy ID:

10

Taxonomy GUID:

87fadfc9-dff3-4f56-b96b-1b95737e5e25

Description:

Content Filters:

Max Categories:

0 (Default)

Default Threshold:

0

Create Default Clues:

Use Global

Default Clue Score:

0

Default Metadata Clue:

Count Mode:

Total

Option	Description
Content Filters	This field allows the taxonomy to be restricted based on a booleanfilter (e.g. using the “CSE-FOLDERS” field) or any of the 8 documentidfilters. See the associated design guide for more information about the ContentFilter field in the Taxonomies table.
Max Categories	Sets the maximum number of classes from this taxonomy that will be

Option	Description
	allocated to each document. To set the Max Categories value across all taxonomies use the Settings tab in Index Manager.
Default Threshold	Sets the default threshold for newly created terms within the selected taxonomy (does not affect existing terms).
Create Default Clues	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.
Default Clue Score	Sets the default score value for new clues.
Default Metadata Clue	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.
Count Mode	Sets the display mode for counts in the treeview.
Show Empty Nodes	Sets the display mode for empty nodes in the treeview.
Synchronise Termset	Enables/Disables automatic synchronisation through the TermStoreManager tool for the whole Term Set.

Class / Term Level

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

Environment

Source Filter: <https://conceptsearching.sharepoint.c...>

Clues Search Browse Working Set Related Settings Logs

Term Settings

Class ID: 3401

Term GUID: bb2f53a0-bd17-4fe6-95a2-4f5466f9fa1a

Description:

Available for Tagging: ☒

Synchronise Term: ☒

Term Weightings

Relevance Threshold:

Parent Boost:

10

%

☒ Default

Child Boost:

10

%

☒ Default

Sibling Boost:

2

%

☒ Default

Related Boost:

10

%

☒ Default

Grandparent Boost:

5

%

☒ Default

Grandchild Boost:

5

%

☒ Default

No Children Boost:

10

%

☒ Default

Remove Boosts

Use Defaults

Save

Option	Description
Available for Tagging	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.
Synchronise Term	Enables / Disables automatic synchronisation through the TermStoreManager tool for the term and its children.
Relevance Threshold	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).
Boosts	<p>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.</p> <p>It is possible to set the “<i>Child</i>” 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 “<i>England</i>” it should also be tagged as “<i>Europe</i>”.</p>

5.3.2. Labels

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

107/207

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:

Microsoft Office Classification Labels				
Link this term to one or more labels in Microsoft Office Security and Compliance. This will automatically set the label of tagged documents in SharePoint and apply any associated records/retention behaviour defined for the label. If more than one label is specified the first label available at the destination site collection will be used.				
	Delete			
	Add			
<input type="checkbox"/>	Priority ^	Label ↕	Last Synced ↕	Available ↕
<input type="checkbox"/>	1	Secure	2019-03-05 12:08:46	<input checked="" type="checkbox"/>
Copy CSV XLSX Showing 1 record(s) Page Size: 10 25 50 100 200				

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:

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.

Select a taxonomy source
Choose to create or import a taxonomy

Connect to taxonomy source
Specify location and/or credentials for the taxonomy source

Select taxonomies
Choose which taxonomies to create/import

What should the taxonomy be called?
The taxonomy name should be unique among all SQL taxonomies

Demo

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://netrix.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.

The screenshot shows a wizard interface for selecting a taxonomy source. It has three steps: 'Select a taxonomy source' (Choose to create or import a taxonomy), 'Connect to taxonomy source' (Specify location and/or credentials for the taxonomy source), and 'Select taxonomies' (Choose which taxonomies to create/import). The first step is active and shows four options: 'New' (Create a new SQL taxonomy), 'Upload' (Upload a taxonomy via XML), 'SharePoint' (Connect to a SharePoint TermStore), and 'Load' (Load a default taxonomy).

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:

The screenshot shows a row of three links: 'Update | Edit | Delete'. The 'Update' link is highlighted in blue.

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

The screenshot shows a wizard interface for updating a prebuilt taxonomy. It has three steps: 'Upload' (Choose which XML file to use for the updating process), 'Options' (Choose how you would like the taxonomy updated), and 'Summary' (Preview the changes before committing them to the taxonomy). The first step is active and shows a question: 'Update this prebuilt taxonomy from the latest definition? Would you like to merge the current terms and clues from the prebuilt taxonomy?'. There are two radio buttons: 'Yes' and 'No'. The 'No' button is selected. Below this is a section for selecting an exported taxonomy file: 'Select the exported taxonomy file: The file should be in the standard conceptSearching taxonomy XML format'. There is a text input field and a 'Browse...' button.

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.

Upload
Choose which XML file to use for the updating process

>

Options
Choose how you would like the taxonomy updated

>

Summary
Preview the changes before committing them to the taxonomy

Which taxonomy do you want to import?
Select which taxonomy contained within the selected file you would like to use

Sensitive Information

How would you like to perform the merge?
You can choose to only process new and updated terms or you can process all changes (including deletions)

☐ Full (including movements / deletions)
 ☒ Add and update only

How would you like to merge clues?
Optionally choose to retain custom clues not marked as 'Predefined'

☐ Retain custom clues
 ☒ Replace all clues

Create Backup
Would you like to create a backup of this taxonomy before applying the update?

[Download](#)

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

Clue label/reference

Is Predefined?

☒

Save

Cancel

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

☐

* Sensitive Information

38a54183-819b-45b9-84aa-9ff98eab011b

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:

Global Settings

Taxonomies Boosts

Note: Deleting external taxonomy registrations (SharePoint) does not delete the source taxonomy. The only effect is that the taxonomy is de-registered from the conceptSearching environment.

Delete Export Add

<input type="checkbox"/>	Name ^	Group Name ^	Status ^	Location ^	Username ^	Search...
<input type="checkbox"/>	Agriculture e0a28f56-ad97-47c4-8949-c3e1c1a6737c	365 Taxonomies	✓ Online ⓘ	https://conceptsearching.sharepoint.com		Test Edit Delete
<input type="checkbox"/>	File Type #0278af-0ff7-4c37-b569-3d3a44c8919c		✓ Online ⓘ	SQL		Update Edit Delete
<input type="checkbox"/>	IPSV 11ec048f-7fc0-4ce7-9df6-1927ba143ce4	365 Taxonomies	✓ Online ⓘ	https://conceptsearching.sharepoint.com		Test Edit Delete
<input type="checkbox"/>	Language d73878a4-3f80-4966-8ccc-b3802fd2d043		✓ Online ⓘ	SQL		Update Edit Delete
<input type="checkbox"/>	* Sensitive Information 38a54183-819b-45b9-84aa-9ff98eab011b		✓ Online ⓘ	SQL		Update Edit Delete
<input type="checkbox"/>	Test 13125d28-2bd8-4806-8c68-ad9dc383d50d		✓ Online ⓘ	SQL		Update Edit Delete

Copy | CSV | XLSX Showing 6 record(s) Page Size: 10 | 25 | 50 | 100 | 200

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:

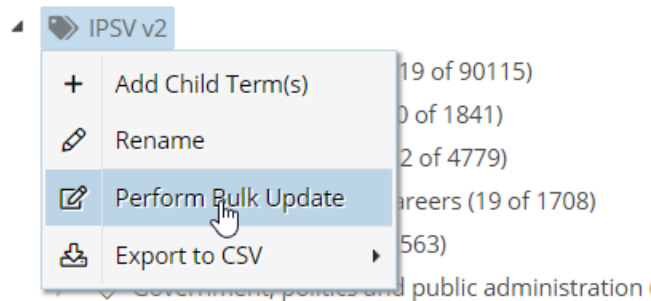
1. On the **Global Settings** tab, go to **Taxonomies**, select the one you need and click **Compare**.
2. In the **Compare** dialog, select what taxonomy definition to compare to:
 - To compare with the latest predefined definition, click **Yes**.
 - Otherwise, click **No** and browse to the required **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:

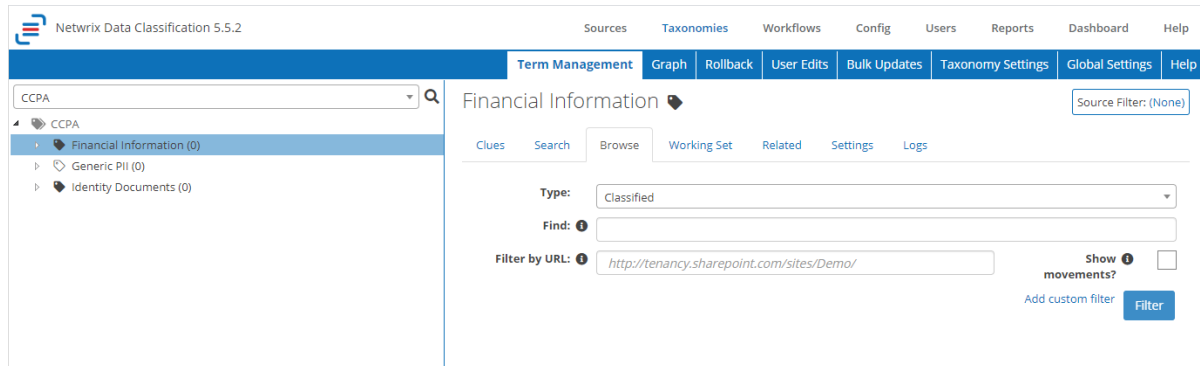
- **Filters**—filters for which terms/clues you wish to update (based on score, clue text, etc).
- **Descendants Limit**—specify how many levels down the update should process within the tree.
- **Exclusions**—specific terms to exclude from the update.

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

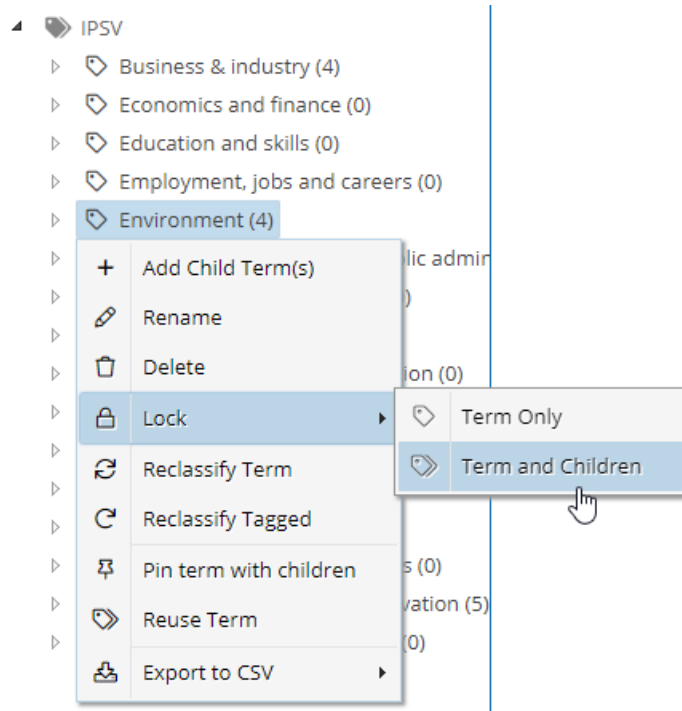
When several users are maintaining the taxonomy structure simultaneously there is a need to prevent concurrent access to individual classes so that one user's work is not overwritten by another user working in the same area of the taxonomy.

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.

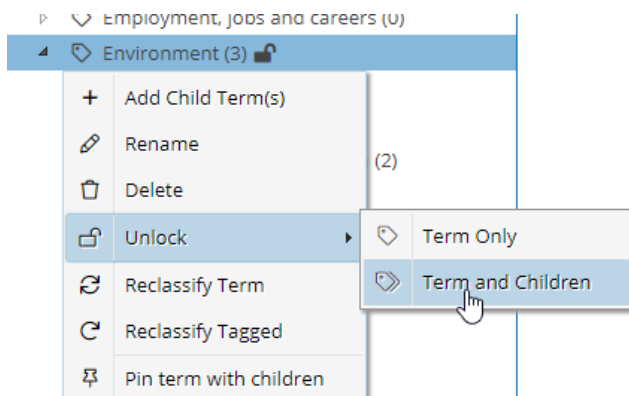
In order to enable this facility the administrator should "Enable User Locking" under **Config** → **Core** → **Query Server**.

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.

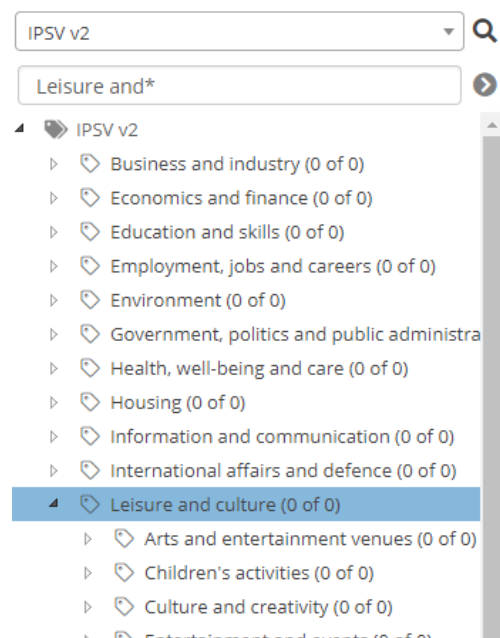
Review the following for additional information:

- [Searching for Taxonomy Terms](#)
- ["Sync Enabled" Treeview Filter](#)
- [Source Filter](#)

Searching for Taxonomy Terms

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

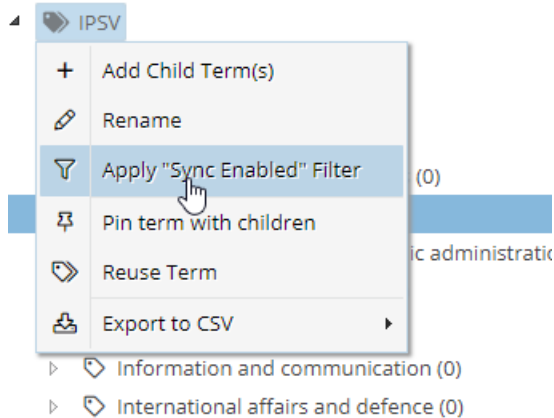
Click the magnifying glass icon to the right of the taxonomy dropdown and a new edit box appears where search text may be entered:



"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:



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://conceptsearching.sharepoint.c...>

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

5.7. Classification Rules (Clues)

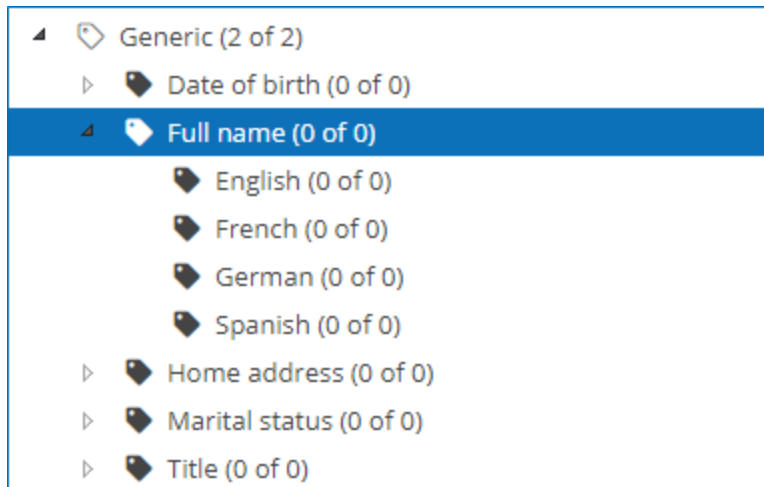
Each taxonomy contains a set of terms. **Terms** are defined by set of configuration **rules** (also called **clues**). Clues are used to describe the language found in documents, making these documents belong to a particular topic.

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:

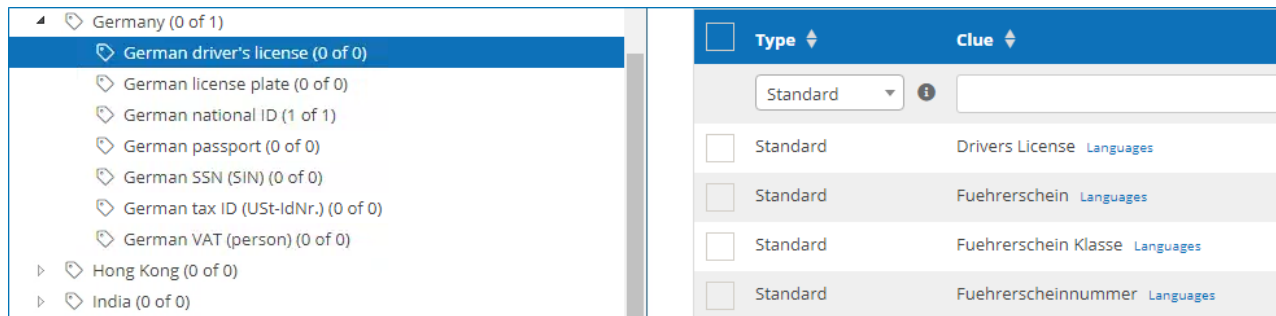
- English
- French
- German

- Spanish



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

- 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:

The screenshot displays the 'Environment' taxonomy page. On the left, a taxonomy tree shows various categories, with 'Environment (3)' selected. The main area shows a table of clues under the 'Clues' tab. The table has columns for Type, Clue, #, Score, and actions (Edit, Delete). The top row is for adding a new clue, and the following rows list existing clues related to natural resources, environmental protection, climate, pollution, conservation, protection, wildlife, and the environment itself.


Type	Clue	#	Score	Actions
Standard	Natural resources Languages	8	40	Edit Delete
Standard	environmental protection Languages	8	30	Edit Delete
Standard	climate Languages	8	20	Edit Delete
Standard	pollution Languages	8	20	Edit Delete
Standard	conservation Languages	8	10	Edit Delete
Standard	protection Languages	8	10	Edit Delete
Standard	Wildlife Languages	8	10	Edit Delete
Standard	Environment Languages	8	10	Edit Delete

Showing 8 record(s) | Page Size: 10 | 25 | 50 | 100 | 200

- 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:

Environment 


Source Filter: (None)

Clues Search Browse Working Set Graph Settings Logs

Delete Search Copy/Move
Bulk Insert Bulk Edit Doc Counts Suggest Clues

<input type="checkbox"/>	Type 	Clue 	#	Score 	*	<input type="text" value="Search..."/>
<input type="checkbox"/>	Standard	<input type="text" value="Natural resources Languages"/>	 -	50		<input type="checkbox"/> Insert
<input type="checkbox"/>	Standard	Natural resources Languages		95	40	 Edit Delete
<input type="checkbox"/>	Standard	environmental protection Languages		52	30	 Edit Delete
<input type="checkbox"/>	Standard	climate Languages		313	20	 Edit Delete
<input type="checkbox"/>	Standard	pollution Languages		104	20	 Edit Delete
<input type="checkbox"/>	Standard	conservation Languages		497	10	 Edit Delete
<input type="checkbox"/>	Standard	protection Languages		865	10	 Edit Delete
<input type="checkbox"/>	Standard	Wildlife Languages		147	10	 Edit Delete
<input type="checkbox"/>	Standard	Environment Languages		535	10	 Edit Delete

Showing 8 record(s) Page Size: 10 | 25 | 50 | 100 | 200

 Info
Document counts loaded

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

- 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 sm?th*
- 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): # 📅 🔍

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

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/*

```
text/html; charset=utf-8
```

```
pdf
```

```
application/pdf
```

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

- Regex-SSN:407-54-8831
- Regex-SSN:407-54-8832

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).

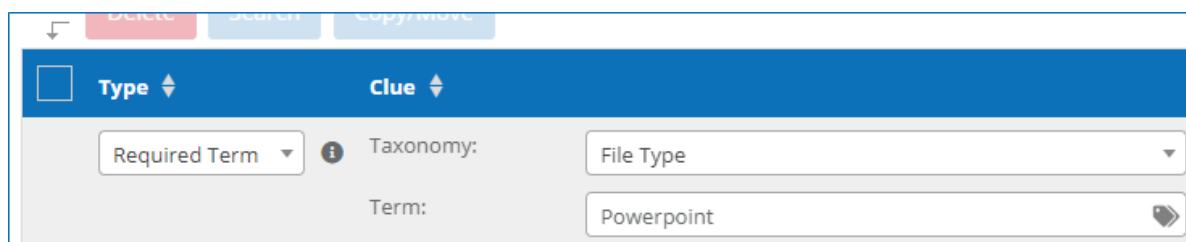
Required Terms clue

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:

The screenshot shows a web-based configuration interface for a 'Required Term' clue. At the top, there are three buttons: 'Delete' (red), 'Search' (blue), and 'Copy/move' (blue). Below these is a blue header bar with a square icon, the word 'Type' with a double-headed arrow, and the word 'Clue' with a double-headed arrow. The main area has a light gray background. On the left, there is a dropdown menu labeled 'Required Term' with a downward arrow. To its right is an information icon (i) and the label 'Taxonomy:'. Further right is a dropdown menu showing 'File Type'. Below the 'Taxonomy' label is the label 'Term:', followed by a text input field containing 'Powerpoint' and a right-pointing arrow icon.

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 nodes 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.

Environment

Source Filter: (None)

Clues

Search

Browse

Working Set

Graph

Settings

Logs

Delete

Search

Copy/Move

Bulk Insert

Bulk Edit

Doc Counts

Suggest Clues

<input type="checkbox"/>	Type	Clue	#	Score	*	Search...
<input type="checkbox"/>	Standard	<input type="text"/>	-	50	<input type="checkbox"/>	Insert
<input type="checkbox"/>	Standard	Natural resources Languages	8	95	40	Edit Delete
<input type="checkbox"/>	Standard	environmental protection Languages	8	52	30	Edit Delete
<input type="checkbox"/>	Standard	climate Languages	8	313	20	Edit Delete
<input type="checkbox"/>	Standard	pollution Languages	8	104	20	Edit Delete
<input type="checkbox"/>	Standard	conservation Languages	8	497	10	Edit Delete
<input type="checkbox"/>	Standard	protection Languages	8	865	10	Edit Delete
<input type="checkbox"/>	Standard	Wildlife Languages	8	147	10	Edit Delete
<input type="checkbox"/>	Standard	Environment Languages	8	535	10	Edit Delete

Showing 8 record(s)

Page Size: 10 | 25 | 50 | 100 | 200

Info

Document counts loaded

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

- 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.

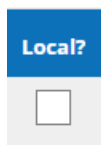
The mandatory clue selector is denoted by the * icon:



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).



- 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:

Type	Clue	Score	*
Standard	Natural resources	40	<input type="checkbox"/>
Standard	environmental protection	30	<input type="checkbox"/>
Standard	climate	20	<input type="checkbox"/>
Standard	pollution	20	<input type="checkbox"/>
Standard	conservation	10	<input type="checkbox"/>
Standard	protection	10	<input type="checkbox"/>
Standard	Wildlife	10	<input type="checkbox"/>
Standard	Environment	10	<input type="checkbox"/>

When this link is used the form changes into a grid editor and many values can be changes 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:

Details			×
Detected Changes			
Clue	Type	Change	
Natural resources	Standard	Score decrease of -20 for ~4 docs	
environmental protection	Standard	Score increase of 5 for ~3 docs	
climate	Standard	Score increase of 2 for ~2 docs	
Skipped			
Clue	Type	Reason	
pollution	Standard	No difference detected	
conservation	Standard	No difference detected	
protection	Standard	No difference detected	
Wildlife	Standard	No difference detected	
Environment	Standard	No difference detected	
			Cancel

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:

Insert

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

Type	Clue	Score
Standard		50
Standard		50
Standard		50
Standard		50
Standard		50
Standard		50
Standard		50

Insert Cancel

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.

Environment

Source Filter: <https://2010.conceptsearching.com>

Clues Search Browse Working Set Graph Settings Logs

Find: energy and fuel

Filter by URL: <http://tenancy.sharepoint.com/sites/Demo/> Show movements? ☐

Add custom filter Search

Suggest Clues Add to Working Set Add to Negative Working Set Re-Index Re-Classify

Showing 7 of 7 record(s) Suggest clues for Search

1 https://2010.conceptsearching.com/Shared Documents/ERD_Safety_Guideline_R2.pdf (100%)

• Whenever possible, reduce operating systems to a zero **energy** state, that is, release all pressure from hydraulic, drilling fluid and air pressure systems, prior to performing maintenance. Use extreme caution when opening drain plugs, pressure caps, valves, and removing hoses and hydraulic lines. • Never weld or cut on or near a **fuel** tank. • Replace all caps, plugs, clamps, cables and guards prior to returning the rig to service. • Never modify any part of the mast without permission from the equipment shop. • If it should become necessary to drain oil, **fuel**, hydraulic fluid or any other industrial fluid in the field, never allow the fluid to drain onto the ground.

+ [1156KB] https://2010.conceptsearching.com/Shared Documents/ERD_Safety_Guideline_R2.pdf

2 <https://2010.conceptsearching.com/Shared Documents/02975.pdf> (100%)

, oil fields generate waste hydrocarbons such as "dirty diesel" **fuel** contaminated from pressure testing pipelines. Hydrocarbon waste is currently assessed for significant organic halides (which might poison refinery catalysts), filtered, commingled and processed with the crude oil stream from a field, and then sold to a pipeline. Upon arrival at the refinery, the waste **fuel**, in solution with the crude oil, is refined to useful products. This practice would benefit from increased public education and acceptance, simplified recycling regulations, and a review of liquid oil field wastes acceptable for recycling with little or no treatment.

+ [56KB] <https://2010.conceptsearching.com/Shared Documents/02975.pdf>

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 \(Clues\)](#) 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

Environment

Source Filter: <https://2010.conce...>

Clues Search Browse Working Set Graph Settings Logs

Type: Classified

Find:

Filter by URL: Show movements? ☒

Add custom filter

☐ Showing 4 of 4 record(s) Suggest clues for Search

1 <https://2010.conceptsearching.com/Shared Documen...>

Document score will increase from 159 to 175

PO Box 19662 Anchorage, AK 99519-6612 Stephen Rae Manager, KIC Laboratory Kugaruk Industrial Center Pouch 340065 Prudhoe Bay, AK 99519-6612 Thomas Redmond Environmental Safety and Health Manager Cameo, Inc. PO Box 91890 Anchorage, AK 99509 Bonnie R obinson Geologist US EPA Office of Solid Waste 401 M St SW (OS-323W) Washington, DC 20460 Marianne See [\[Pollution\]](#) Prevention Specialist [\[Pollution\]](#) Prevention Office Alaska Department of Environmental [\[Conservation\]](#) 3601 C St. #1334 Anchorage, AK 99503 Doug Segar Director Environmental and Natural Resources Institute University of Alaska, Anchorage Anchorage, AK 99503 Ben Shafsky Assistant Operations Manager Doyon Drilling, Inc.

[56KB] <https://2010.conceptsearching.com/Shared Docu...>

2 <https://2010.conceptsearching.com/Shared Documen...>

Document score will increase from 159 to 175

PO Box 19662 Anchorage, AK 99519-6612 Stephen Rae Manager, KIC Laboratory Kugaruk Industrial Center Pouch 340065 Prudhoe Bay, AK 99519-6612 Thomas Redmond Environmental Safety and Health Manager Cameo, Inc. PO Box 91890 Anchorage, AK 99509 Bonnie R obinson Geologist US EPA Office of Solid Waste 401 M St SW (OS-323W) Washington, DC 20460 Marianne See [\[Pollution\]](#) Prevention Speci

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 \(Clues\)](#) 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.

Environment

Source Filter: (None)

Clues

Search

Browse

Working Set

Graph

Settings

Logs

Type: Classified

Find:

Filter by URL: http://tenancy.sharepoint.com/sites/Demo/

Show movements?

Add custom filter

Filter

Suggest Clues

Add to Working Set

Add to Negative Working Set

Re-Index

Re-Classify

Showing 4 of 4 record(s)

Suggest clues for Search

1

https://2010.conceptsearching.com/Shared Documents/Test/02975.pdf (100%)

PO Box 19662 Anchorage, AK 99519-6612 Stephen Rae Manager, KIC Laboratory Kuparuk Industrial Center Pouch 340065 Prudhoe Bay, AK 99519-6612 Thomas Redmond Environmental Safety and Health Manager Cameo, Inc. PO Box 91890 Anchorage, AK 99509 Bonnie Robinson Geologist US EPA Office of Solid Waste 401 M St SW (OS-323W) Washington, DC 20460 Marianne See Pollution Prevention Specialist Pollution Prevention Office Alaska Department of Environmental Conservation 3601 C St. #1334 Anchorage, AK 99503 Doug Segar Director Environmental and Natural Resources Institute University of Alaska, Anchorage Anchorage, AK 99503 Ben Shafsky Assistant Operations Manager Doyon Drilling, Inc.

+

[56KB] https://2010.conceptsearching.com/Shared Documents/Test/02975.pdf

2

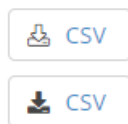
https://2010.conceptsearching.com/Shared Documents/02975.pdf (100%)

PO Box 19662 Anchorage, AK 99519-6612 Stephen Rae Manager, KIC Laboratory Kuparuk Industrial Center Pouch 340065 Prudhoe Bay, AK 99519-6612 Thomas Redmond Environmental Safety and Health Manager Cameo, Inc. PO Box 91890 Anchorage, AK 99509 Bonnie Robinson Geologist US EPA Office of Solid Waste 401 M St SW (OS-323W) Washington, DC 20460 Marianne See Pollution Prevention Specialist Pollution Prevention Office Alaska Department of Environmental Conservation 3601 C St. #1334 Anchorage, AK

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:



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:

Export

Notification Email Group:

None

Quantity to Extract:

All results

Export

Cancel

5.8. Suggestions

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

Type	Clue	Score	*
Standard	Natural resources	40	<input type="checkbox"/>
Standard	environmental protection	30	<input type="checkbox"/>
Standard	climate	20	<input type="checkbox"/>
Standard	pollution	20	<input type="checkbox"/>
Standard	conservation	10	<input type="checkbox"/>
Standard	protection	10	<input type="checkbox"/>
Standard	Wildlife	10	<input type="checkbox"/>
Standard	Environment	10	<input type="checkbox"/>

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:

Environment

Source Filter: (None)

Clues Search Browse Working Set Graph Settings Logs

Type: Classified

Find:

Filter by URL: Show movements? ☐

Add custom filter

☐ Showing 4 of 4 record(s) Suggest clues for Search

1	https://2010.conceptsearching.com/Shared Documents/Test/02975.pdf (100%) <div> </div> <p>PO Box 19662 Anchorage, AK 99519-6612 Stephen Rae Manager, KIC Laboratory Kuparuk Industrial Center Pouch 340065 Prudhoe Bay, AK 99519-6612 Thomas Redmond Environmental Safety and Health Manager Cameo, Inc. PO Box 91890 Anchorage, AK 99509 Bonnie Robinson Geologist US EPA Office of Solid Waste 401 M St SW (OS-323W) Washington, DC 20460 Marianne See Pollution Prevention Specialist Pollution Prevention Office Alaska Department of Environmental Conservation 3601 C St. #1334 Anchorage, AK 99503 Doug Segar Director Environmental and Natural Resources Institute University of Alaska, Anchorage Anchorage, AK 99503 Ben Shafsky Assistant Operations Manager Doyon Drilling, Inc.</p> <p> [56KB] https://2010.conceptsearching.com/Shared Documents/Test/02975.pdf</p>
2	https://2010.conceptsearching.com/Shared Documents/02975.pdf (100%) <div> </div> <p>PO Box 19662 Anchorage, AK 99519-6612 Stephen Rae Manager, KIC Laboratory Kuparuk Industrial Center Pouch 340065 Prudhoe Bay, AK 99519-6612 Thomas Redmond Environmental Safety and Health Manager Cameo, Inc. PO Box 91890 Anchorage, AK 99509 Bonnie Robinson Geologist US EPA Office of Solid Waste 401 M St SW (OS-323W) Washington, DC 20460 Marianne See Pollution Prevention Specialist Pollution Prevention Office Alaska Department of Environmental Conservation 3601 C St. #1334 Anchorage, AK</p>

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.

Breton

Source Filter: (None)

Clues Search Browse Working Set **Related** Graph Settings Logs

Related Terms

Delete Add

Related From	Related To	
<input type="checkbox"/> Breton	Basque	<input type="text" value="Search..."/> Delete

Copy | CSV | XLSX Showing 1 record(s) Page Size: 10 | 25 | 50 | 100 | 200

Term Locations

Term Locations

Language > Breton

Copy | CSV | XLSX Showing 1 record(s) Page Size: 10 | 25 | 50 | 100 | 200

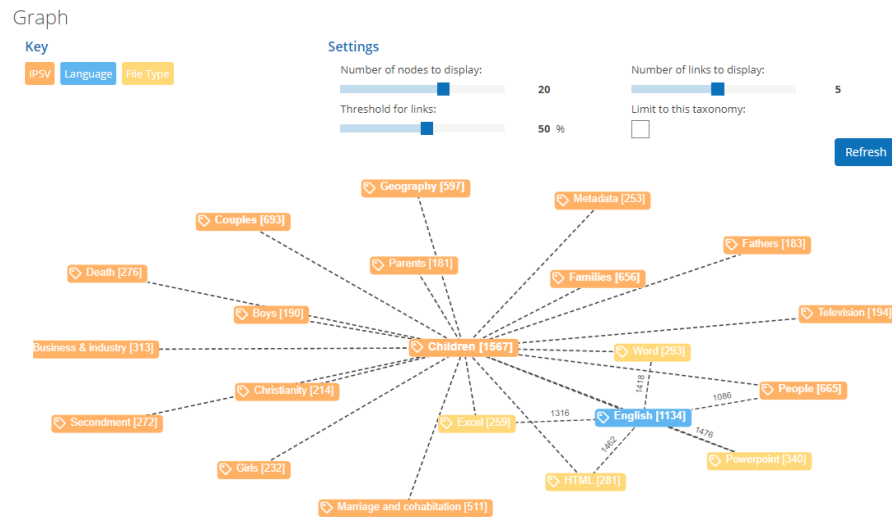
When a term is located in multiple branches of the taxonomy (a polyhierarchical taxonomy) – the **Related** tab will also display each of the locations to allow you to jump to the specific branch.

5.11. Additional Configuration

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

Tab	Description
Graph	The Graph tab shows a graphical representation of classification intersection points.

Tab	Description
-----	-------------



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:
------	--

Environment

Source Filter: <https://conceptsearchin...>

Clues Search Browse Working Set Related Settings **Logs**

[Clear Logs](#)

Date / Time	Description	Username
2019-03-05 12:55:49	Insert Clue ("environment") for Term: Environment	
2019-03-05 12:55:43	Delete Clue ("Environment") from Term: Environment	

[Copy](#) | [CSV](#) | [XLSX](#) | Showing 2 record(s) | Page Size: **10** | 25 | 50 | 100 | 200


User Edits	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:
------------	---

Tab

Description

User Edits

Class Name ^	User Adds ⬆	User Deletes ⬆	Total User Edits ⬆	<input type="text" value="Search..."/>
No records to display				

 [Copy](#) | [CSV](#) | [XLSX](#)

Showing 0 record(s)


Page Size: [10](#) | [25](#) | [50](#) | [100](#) | [200](#)

User Suggestions An optional interface can be enabled to allow users to suggest new terms for the termset hierarchy
(<http://netwrixdataclassificationserver/conceptQS/Taxonomies/TermSuggest.aspx>).
Suggestions can trigger automatic notifications to taxonomy administrators, as well as being recorded in the database for later review on the "User Suggestions" tab:

User Suggestions

The following are suggestions for terms to be added to the termset hierarchy:

Display: All Pending Accepted Rejected									Add
Term ^	Suggested Path	Requestor	Email Address	Reason	Request Date	Processed Date	Processed By	Status	Search...
DICOM	Root	John Smith	john@demo.com	Identification of DICOM files	2019-03-05 13:04:40	-	-	-	Accept Reject

 Copy | CSV | XLSX

Showing 1 record(s)

Page Size: 10 | 25 | 50 | 100 | 200

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		Add	
Name ^	Workflows ^	Search...	
Global	Demo, Quarantine	Delete	
Global for SharePoint	Migrate	Delete	
		Copy CSV XLSX	
		Showing 2 record(s)	
		Page Size: 10 25 50 100 200	

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):

Workflows			
Workflows			Logs
Workflows			Add
Name	Workflows	Search...	
Global	My Workflow		Delete
Global for Box	Box Test		Delete
Global for Exchange	TestForExchange		Delete
Global for File	TestForAll		Delete
Global for Google Drive	Google Drive workflow		Delete
Global for SharePoint	Email on SharePoint content (not HTML or XML), SharePoint test		Delete

Copy | CSV | XLSX Showing 6 record(s) Page Size: 10 | 25 | 50 | 100 | 200

NFR QC TEST | Netwrix Demo | FMWSAPERF\Administrator © 2019 Netwrix Corporation netwrix

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).

The screenshot shows the 'Workflows' management page in the Netwrix Data Classification application. The page has a blue header with tabs for 'Workflows', 'Configs', 'Plugins', and 'Logs'. Below the header, the 'Workflows' section is active, showing a breadcrumb 'Workflows > Global for Google Drive'. There are buttons for 'Delete', 'Enable', 'Disable', and 'Add'. A table lists the workflows with columns: 'Name', 'Details', 'Rules', and 'Active'. One workflow, 'Google Drive workflow', is listed with 'Email Alert' details, 1 rule, and is active (indicated by a green checkmark). Action links 'Disable', 'Clone', 'Rename', and 'Delete' are available for this workflow. Below the table, there are links for 'Copy', 'CSV', and 'XLSX', and a status 'Showing 1 record(s)'. A 'Page Size' dropdown is set to '25'. The footer contains the text 'NFR QC TEST | Netwrix Demo | FMWSAPERF\Administrator', '© 2019 Netwrix Corporation', and the Netwrix logo.

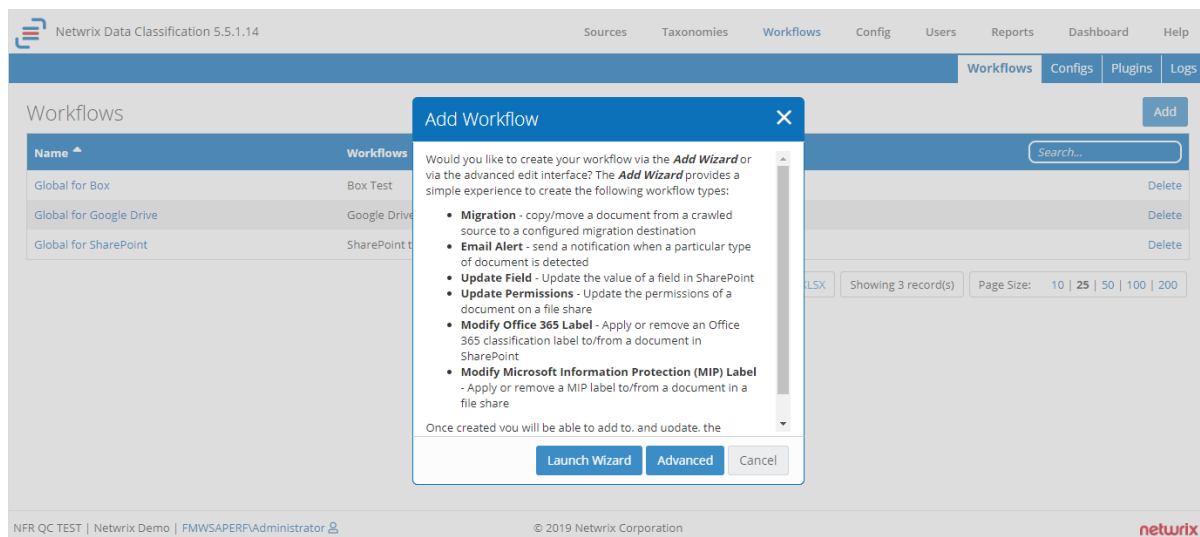
Name	Details	Rules	Active
Google Drive workflow	Email Alert	1	✓

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.

The screenshot shows the 'Add Workflow' wizard interface. The title bar is blue with a close button. Below the title bar is a progress bar with four steps: 'Which content source(s)?' (selected), 'What do you want to do?', 'When do you want to do it?', and 'Summary'. The main content area has a heading 'What sort of documents would you like this workflow to target?' followed by a subtext 'You can either target all documents of a particular type (such as all documents stored in SharePoint) or, target a particular source.' Below this is a dropdown menu showing 'Google Drive'. Another heading 'Which source(s) would you like this workflow to target?' is followed by a subtext 'You can optionally restrict the workflow to only execute against a single source (or, group of sources)'. Below this is a dropdown menu showing 'All sources'. At the bottom right are 'Next' and 'Cancel' buttons.

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](#).


Add Workflow

 Which content source(s)?
The document type(s) to target


>


 What do you want to do?
Choose the action(s) to carry out

>

 When do you want to do it?
Restrict when the workflow runs

>

 Summary
Review the configuration

 **Email Alert**
Send an email alert to a static email address.

Back

Cancel

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 [X]

Which content source(s)? > What do you want to do? > **When do you want to do it?** > Summary

Which documents do you want this workflow to target?

You can choose to run this workflow against:

- All documents in the selected source(s), or
- Documents that have been tagged with any classifications, or
- Document with specific classifications, or
- Documents that have not been tagged with specific classifications

☐ Any Document ☐ Any Classified Document ☒ Specific Classifications

Identify specific documents by choosing one or more terms that a document should either be tagged with ("Classified") or not tagged with ("Not Classified")

☒ Classified ☐ Not Classified

File Type > Text

[Back] [Next] [Cancel]

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

Add Workflow [X]

Which content source(s)? > What do you want to do? > **When do you want to do it?** > Summary

Which documents do you want this workflow to target?

You can choose to run this workflow against:

- All documents in the selected source(s), or
- Documents that have been tagged with any classifications, or
- Document with specific classifications, or
- Documents that have not been tagged with specific classifications

☐ Any Document ☐ Any Classified Document ☒ Specific Classifications

Identify specific documents by choosing one or more terms that a document should either be tagged with ("Classified") or not tagged with ("Not Classified")

☐ Classified ☒ Not Classified

File Type > Excel **File Type > Powerpoint**

Require all conditions?

Do you require the document to match one of the specified conditions - or all of the specified conditions?

☒ All ☐ Any

[Back] [Next] [Cancel]

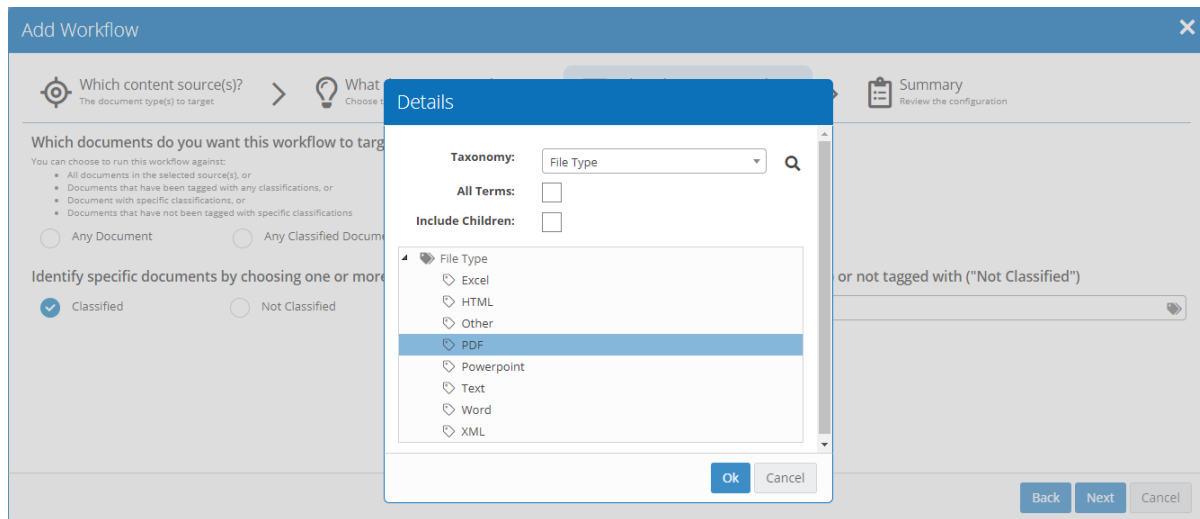
If you have selected more than one filter, you will be prompted what logic should be used when applying the filters:

- 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.

Add Workflow

Which content source(s)?
The document type(s) to target

What do you want to do?
Choose the action(s) to carry out

When do you want to do it?
Restrict when the workflow runs

Summary
Review the configuration

Which documents do you want this workflow to target?

You can choose to run this workflow against:

- All documents in the selected source(s), or
- Documents that have been tagged with any classifications, or
- Documents with specific classifications, or
- Documents that have not been tagged with specific classifications

☐ Any Document☐ Any Classified Document☒ Specific Classifications

Identify specific documents by choosing one or more terms that a document should either be tagged with ("Classified") or not tagged with ("Not Classified")

☐ Classified☒ Not Classified

Select Term

✕ File Type > HTML

✕ File Type > XML

Require all conditions?

Do you require the document to match one of the specified conditions - or all of the specified conditions?

☐ All☒ Any

Back

Next

Cancel

154/207

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

Which content source(s)?
The document type(s) to target

What do you want to do?
Choose the action(s) to carry out

When do you want to do it?
Restrict when the workflow runs

Summary
Review the configuration

Choose a name for your workflow
The name should be used to uniquely identify the functionality of the Workflow at a high level. You can select any name more than 3 characters in length.

Should this workflow be enabled on creation?
Would you like documents to begin being processed by this workflow immediately? Documents that have already been classified will need to be re-classified for the workflow to execute.

☐ Enabled
☒ Disabled

Which content source(s)?

Source Type: SharePoint
Sources: All sources

What do you want to do?

Action: Send an Email
Recipient(s): spadmin@acme.com
Send From: ndc@acme.com (172.16.6.35:465)
Subject: Workflow Rule run on: [cs:PageTitle]
Body: A rule has been run on the following document: [cs:PageTitle] Classifications: [cs:Classifications]

When do you want to do it?

Run this workflow against : Documents with Specific Classifications
Not classified as:

- File Type > HTML, or
- File Type > XML

Back Add Cancel

When finished, click **Add** to close the wizard. Your new workflow will be added to the list on the **Workflows** tab:

Workflows			Workflows	Configs	Plugins	Logs
Workflows			Add			
Name	Workflows		Search...			
Global for Box	Box Test		Delete			
Global for Google Drive	Google Drive workflow		Delete			
Global for SharePoint	Email on SharePoint content (not HTML or XML)	SharePoint test	Delete			
			Copy CSV XLSX			
			Showing 3 record(s)			
			Page Size: 10 25 50 100 200			

NFR QC TEST | Netwrix Demo | FMWSAPERF\Administrator

© 2019 Netwrix Corporation

netwrix

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.

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.

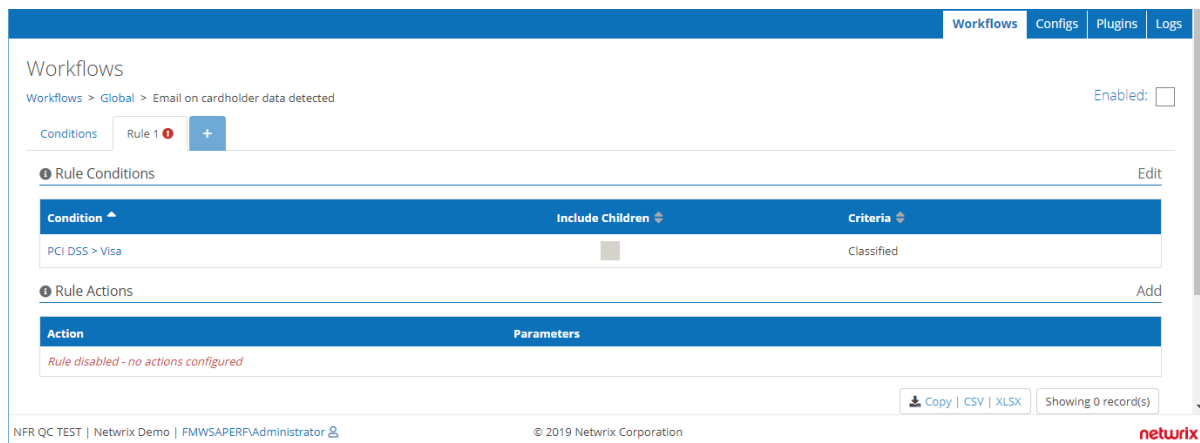
3. Finally, click **OK** to save the settings and close the dialog.

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**.

The 'Details' window has a blue header. Below it, there is a 'Taxonomy:' dropdown menu with 'PCI DSS' selected and a search icon. Below this are two checkboxes: 'All Terms:' and 'Include Children:', both of which are unchecked. A tree view shows the hierarchy under 'PCI DSS', including 'AMEX', 'Diners Club', 'Discover', 'Generic PCI DSS', 'JCB', 'Mastercard', 'UnionPay', and 'Visa'. The 'Visa' item is highlighted with a blue background. At the bottom right, there are 'Ok' and 'Cancel' buttons.

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**.


The 'Add Action' dialog box has a blue header with a close button. It contains several fields: 'Action Type:' with a dropdown menu showing 'Email Alert'; a description 'Send an email alert to a static email address.'; 'Email Address' with a text field containing 'securitydept@acme.com'; 'SMTP Config' with a dropdown menu showing '172.16.6.35:465 (ndc@acme.com)'; 'Subject' with a text field containing 'Workflow Rule run on: [cs:PageTitle]'; and 'Email Body Template' with a text area containing 'A rule has been run on the following document:' and '[cs:PageTitle]'. At the bottom right, there are 'Save' and 'Cancel' buttons.

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.

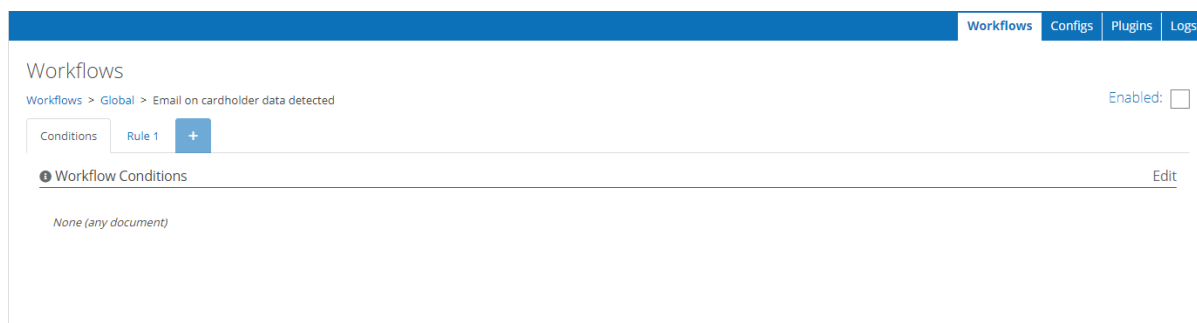
If multiple rule actions have been configured, they will be processed in the order listed. Use the red down arrow or green up-arrow to change the processing sequence as required:

y=Documents, e=, tem=false,	↓ Add Migration Action Edit Delete	
n:587 Title], EmailBody=A rule has been run	⬇ Edit Delete	
	⬆ Edit Delete	
	⬆ Edit Delete	
<div>  Copy CSV XLSX </div> <div>Showing 4 record(s)</div>		

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.

Workflows

Workflows > Global > Email on cardholder data detected

Enabled: ☐

Conditions Rule 1 +

Rule Conditions Edit

Condition	Include Children	Criteria
PCI DSS > Visa		Classified

Rule Actions Add

Action	Parameters
Rule disabled - no actions configured	

Copy CSV XLSX Showing 0 record(s)

NFR QC TEST | Netwrix Demo | FMWSAPERF\Administrator © 2019 Netwrix Corporation netwrix

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):

The screenshot shows the 'Workflows' section of the Netwrix interface. At the top, there are tabs for 'Workflows', 'Configs', 'Plugins', and 'Logs'. Below the tabs, there's a search bar and an 'Add' button. The main table has columns for 'Name' and 'Workflows'. The 'Name' column lists various scopes: 'Global', 'Global for Box', 'Global for Exchange', 'Global for File', 'Global for Google Drive' (highlighted with a red box), and 'Global for SharePoint'. The 'Workflows' column lists the corresponding workflows: 'My Workflow', 'Box Test', 'TestForExchange', 'TestForAll', 'Google Drive workflow', and 'Email on SharePoint content (not HTML or XML), SharePoint test'. Each row has a 'Delete' button on the right. At the bottom, there are links for 'Copy', 'CSV', and 'XLSX', and a status bar showing 'Showing 6 record(s)' and 'Page Size: 10 | 25 | 50 | 100 | 200'.

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

The screenshot shows the 'Workflows' section for the 'Global for Google Drive' scope. At the top, there are tabs for 'Workflows', 'Configs', 'Plugins', and 'Logs'. Below the tabs, there's a search bar and an 'Add' button. The main table has columns for 'Name', 'Details', 'Rules', and 'Active'. The 'Name' column lists the workflow: 'Google Drive workflow'. The 'Details' column shows 'Email Alert'. The 'Rules' column shows '1'. The 'Active' column shows a green checkmark. Each row has a 'Delete' button on the right. At the bottom, there are links for 'Copy', 'CSV', and 'XLSX', and a status bar showing 'Showing 1 record(s)' and 'Page Size: 10 | 25 | 50 | 100 | 200'.

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.

Content source type	Available actions
Exchange	Email Alert Migrate Document Apply Additional Classification Advanced Actions for Exchange* : delete email, move email
File System	Email Alert Migrate Document Apply Additional Classification Advanced Actions for File System* : update permissions, add/remove MIP label
Google Drive	Email Alert Migrate Document Apply Additional Classification
SharePoint	Email Alert Migrate Document Apply Additional Classification

Content source type	Available actions
---------------------	-------------------

[Advanced Actions for SharePoint](#)*: send classification value, filtered targeted meta update, write/remove O365 label, copy/move document

SQL and other databases

[Email Alert](#)

[Migrate Document](#)

[Apply Additional Classification](#)

* — 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.

The screenshot shows the 'Add Workflow' dialog window with the following configuration steps:

- Which content source(s)?**: The document(s) to target.
- What do you want to do?**: Choose the action(s) to carry out. (Selected: Email Alert)
- When do you want to do it?**: Restrict when the workflow runs.
- Summary**: Review the configuration.

Specific recipient(s): Enter one or more recipients. (Field contains: administrator@corp.local)

Who should the email be sent from?: Choose the email account you wish for this email to be sent from. (Field contains: cs@corp.com (172.17.6.35:465))

Email Subject: Define the subject of the email. Certain dynamic parameters can be specified from the crawled content; these can be selected from the associated dropdown list. (Field contains: Workflow Rule run on: [cs:PageTitle])

Email Body Template: Define the content of the email body. Certain dynamic parameters can be specified from the crawled content; these can be selected from the associated dropdown list. (Field contains: A rule has been run on the following document: [cs:PageTitle], Classifications: [cs:Classifications])

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:

Field	Settings to specify
Specific recipients	Specify email address to send the alert to. To enter multiple recipient, click + on the right.
Who should the email be sent from?	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:

Field	Settings to specify
	<ul style="list-style-type: none"> • 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

Email Server Details

Host:

172.17.6.35

Port:

465

☒ Use SSL

From Email:

cs@corp.com

Username:

cs

Password:

.....

Test Configuration Settings

We recommend you test your configuration by entering a confirmation email address below.

Email Address:

ittest@corp.com

Save

Cancel

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

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

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:

Field	Setting to specify
Email Address	<p>Specify email recipients. You can enter multiple static email addresses.</p> <p>NOTE: Dynamic configurations will use the '<i>Document Modified/Created By</i>' metadata value, looking up the user's email address from Active Directory where appropriate.</p>
SMTP Config	<p>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.</p>

Field	Setting to specify
Subject	<p>Specify the template for email subject.</p> <p>The template can contain dynamic values that will be obtained from the crawled content (e.g. <code>[cs:PageUrl]</code>).</p> <p>TIP: To get the list of available fields, click the details link.</p>
Email Body Template	<p>Specify the template for email body.</p> <p>The template can contain dynamic values that will be obtained from the crawled content (e.g. <code>[cs:PageUrl]</code>).</p> <p>TIP: To get the list of available fields, click the details link.</p>

6.3.1.2. Migrate Document

This action can be used to copy or move a document between content sources (from '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.

Type	As 'source'	As 'destination'	Migration Config Type	Supports structured migration?	Move?	Update source item?	Mark source 'read-only'?
Exchange	Yes	No		Yes	No	No	No
File System	Yes	Yes	Custom - File Share	Yes	Yes	Yes	Yes
Google Drive	Yes	Yes	Source (Google Drive account)	Yes	No	No	No
SharePoint	Yes	Yes	Custom - SharePoint Site Collection	Yes	Yes	Yes	No
SQL and other databases	Yes	No		Yes	No	No	No

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:

- If you created several sources for migration destinations, select on 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:

Option	Description
Replicate folder structure	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\User@domain.com\Inbox\HR).
Copy or Move the document	Select one of the following: <ul style="list-style-type: none"> • Copy • Move

Option	Description
Mark Source as Read-only	The original item can be marked as read only.
What action should be taken if the document already exists at the destination	Select action to perform: <ul style="list-style-type: none">• Replace• Append
Redact the document	If update of the source item is supported by the source system, then using this option will instruct the program to apply the redaction plan to the source document after its successful migration. 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.

Add Action
↗ ✕

Action Type:

Migrate To File System
▼

Migration Destination

Please Select
▼

Destination Rename Mode

None
▼

Specifies what action to take if a file exists at the destination with the same name

Maintain Folder Structure
☐

Delete Original Item
☐

Save

Cancel

Setting	Description	Comments
Migration Destination	The root destination to migrate to.	Make sure to define the required destination as Migration Config.
Destination Rename Mode	<p>Specifies what action to take if a file exists at the destination with the same name.</p> <ul style="list-style-type: none"> None – overwrite the destination file or issue a “duplicate” error. <p>NOTE: Behaviour depends on the migration destination.</p> <ul style="list-style-type: none"> Append Number - append a numeric counter as a suffix to the file name (e.g. <i>document_2.txt</i>). 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.	<p>Applies if this capability is supported by the source system.</p> <p>For Exchange, the path will also include a folder for the mailbox name (e.g. <code>\\MigrationDestination\User@domain.com\Inbox\HR</code>).</p>

Setting	Description	Comments
Delete Original Item	If selected, the original item will be deleted after it is successfully copied to the destination.	Applies if this capability is supported by the source system.
Mark Original item as Read Only	If selected, the original item will be marked as <i>read-only</i> .	Applies if this capability is supported by the source system.
Redaction Plan	If redaction plans have been configured, specify the redaction plan to be applied to the document. See Redaction .	By default, this will be applied to the document at the destination.
Redact Original	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.	Note that this option is not available when performing a move (deleting the original item).

Source-specific settings

Settings for Google Drive content migration are described below.

Setting	Description	Comments
Destination Folder	The path to migrate the document to relative to the migration destination.	To migrate to the root folder, leave blank. Destination example: <i>Folder/SubFolder/SubFolder2</i>

Settings for SharePoint content migration are described below.

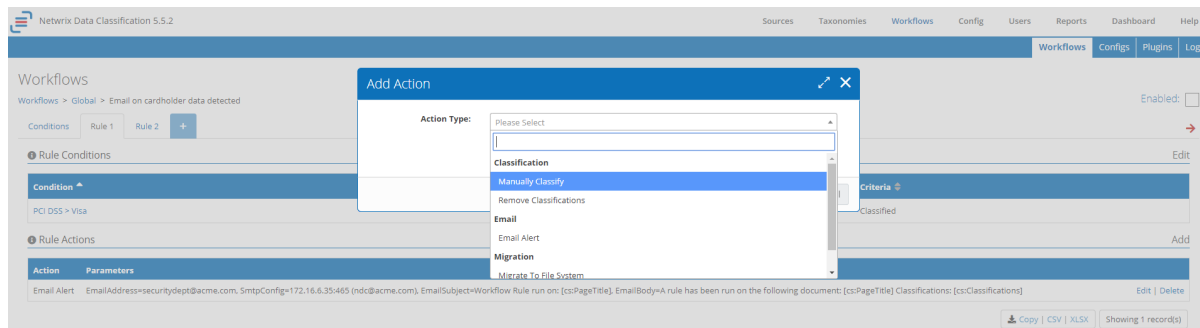
Setting	Description	Comments
Library/Folder	The library and optional subfolder to migrate the document to in the migration destination.	
Mode	Previous versions and metadata can be included.	Only applicable if the source system is also SharePoint.

Setting	Description	Comments
Dynamic Destination Field Name	Specify a metadata field on the item that can be used to dynamically lookup the migration destination.	Only applicable if the source system is also SharePoint.
Web Path	<p>The relative web path for the migration of the document. Format should be any of the following:</p> <ul style="list-style-type: none"> • <i>~/WebPath</i> —a document found, e.g., at <i>http://sharepoint/sites/Test/Demo</i> with the relative path <i>~/Subsite</i> would attempt to migrate to <i>http://sharepoint/sites/Test/Demo/subsite</i> • <i>/SiteCollectionPath</i> —a document found, e.g., at <i>http://sharepoint/sites/Test/Demo</i> with the relative path <i>/Subsite</i> would attempt to migrate to <i>http://sharepoint/sites/Test/Subsite</i> 	Only applicable if a SharePoint relative migration is chosen.
List Title	The name of the library at the web path specified to migrate the document to.	Only applicable if a SharePoint relative migration is chosen.
Fallback - if relative path invalid	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.	Only applicable if a SharePoint relative migration is chosen.

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.

The screenshot shows a dialog window titled "Add Action" with a blue header bar containing a maximize icon and a close button (X). The dialog has a white background and contains the following fields:

- Action Type:** A dropdown menu with "Delete email" selected. Below it is a descriptive text: "Action to remove an email from the Exchange mailbox."
- Delete Mode:** A dropdown menu with "Hard Delete" selected. Below it is a descriptive text: "Matches the internal Microsoft Exchange DeleteMode options."
- Suppress read-receipts:** A checkbox that is checked, indicated by a blue square with a white checkmark.

At the bottom right of the dialog are two buttons: "Save" (blue) and "Cancel" (gray).

Specify the following action parameters:

Action parameter	Description	Comments
Delete Mode	Matches the native Microsoft Exchange Delete Modes : <ul style="list-style-type: none"> • Soft Delete — Email will be available for recovery from the <i>Deleted Items</i> folder. • Hard Delete — Email will not be available for recovery after deletion. • Move to Deleted items — Email will be moved to <i>Deleted Items</i> folder. 	See this Microsoft article for details.
Suppress Read Receipts	With this option selected, <i>Read receipts</i> will not be sent (if requested) for the item being deleted.	Selected by default.

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.

Add Action

Action Type:

Apply MIP Label

Applies a label using Microsoft Information Protection (MIP)

MIP Configuration

Default

The MIP configuration to use for this workflow

Label ID

Applies the specified MIP label to the document.

Justification

The justification for downgrading a label (if applicable)

Save

Cancel

See also [Modify MIP Label](#).

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

Action Type: Send classification value(s)
This action can be used to send classification values to a selected SharePoint property.

Field Name: DocType
Name of the SharePoint field to update.

Taxonomy: Financial Records
The classification values that will be inserted into the specified field

Save **Cancel**

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

Sources Taxonomies Workflows Config Users Reports Dashboard Help

Workflows Configs Plugins Logs

Migration Configs
 Action Configs
 Content Type Hubs

Content Type Hubs

Add

Server URL Username Search...

No records to display

Copy CSV XLSX Showing 0 record(s) Page Size: 10 25 50 100 200

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**

- Once added, navigate back to the main **Workflows** screen, and select the newly added group from the **Workflow Groups** grid
- 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.

Plugins

Plugins will be detected in the following location: C:\Program Files\ConceptSearching\Plugins\

Detect Plugins

Class Name ^	Interface Type ^	Assembly Path ^	Enabled ^	Search...
No records to display				

Copy | CSV | XLSX

Showing 0 record(s)

Page Size: 10 | 25 | 50 | 100 | 200

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.

Logs

Display Period: Past Day | Past Week | Past Month | Past Year | All Time

Clear Logs

Run Date ^	Workflow ^	Action ^	Action Type ^	Page Url ^	Result ^	Search...
No records to display						

Copy | CSV | XLSX

Showing 0 record(s)

Page Size: 10 | 25 | 50 | 100 | 200

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](#)

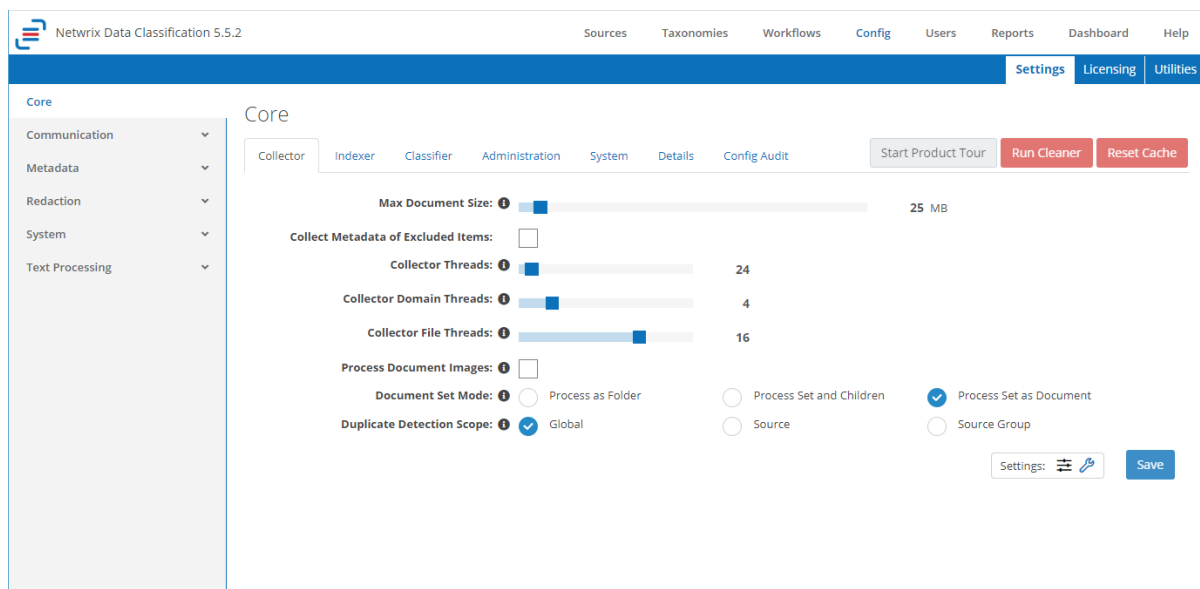
7.1. Index Maintenance

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

1. Open NDC Management Web Console.
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](#)

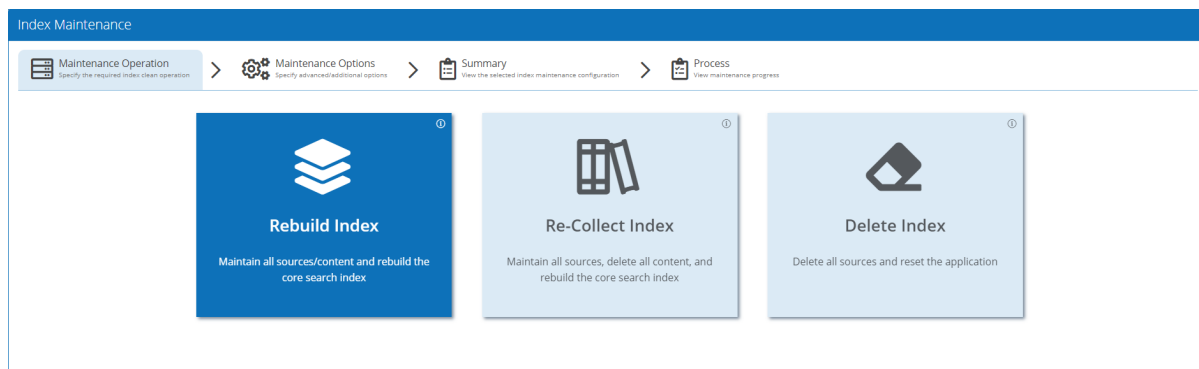
7.1.1. Step 1: Maintenance Operation

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.



7.1.2. Step 2: Maintenance Options

Specify options for the operation you have selected.

Operation selected	Available options	Details
Rebuild Index	Shrink the "text.cse" file? <ul style="list-style-type: none"> ◦ Shrink 	Selecting Shrink will rebuild the <i>Text.cse</i> file, removing any fragmentation. Shrink will require

Operation selected	Available options	Details
	<ul style="list-style-type: none"> Don't Shrink (default) 	sufficient disk space to process (up to the existing size of <i>Text.cse</i>) and may take some time to complete.
All operations	<p>Would you like to re-run the product configuration wizard?</p> <ul style="list-style-type: none"> Run Don't Run (default) 	Select Run if you want to re-configure this instance by going through the initial steps of the product configuration. Note that this will pause all sources.

Index Maintenance

Maintenance Operation > Maintenance Options > **Summary** > Process

Shrink the "text.cse" file?
When selected this will re-build the Text.cse file removing any fragmentation. "Shrink" will require sufficient hard drive space to process (up to the existing size of text.cse) and may take some time to complete.

☐ Shrink ☒ Don't Shrink

Would you like to re-run the product configuration wizard?
Re-configure this instance by going through the initial steps of the product configuration. This will pause all sources.

☐ Run ☒ Don't Run

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.

Index Maintenance

Maintenance Operation > Maintenance Options > **Summary** > Process

Confirm

Are you sure that you wish to rebuild the index?

Summary

Actions: Rebuild index

Run the product configuration wizard: No

Shrink text.cse: No

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.

Core

Communication

Metadata

Redaction

System

Text Processing

Core

Collector

Indexer

Classifier

Query Server

Logging

Details

Start Product Tour

Reset Cache

Max Document Size: 0 MB

Collect Metadata of Excluded Items:

Collector Threads:

0 (auto)

Collector Domain Threads:

0 (auto)

Collector File Threads:

0 (auto)

Process Document Images:

Document Set Mode:

Process as Folder

Process Set and Children

Process Set as Document

Duplicate Detection Scope:

Global

Source

Source Group

Settings:

Save

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 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.

				Settings	Licensing	Utilities
Licensing Summary						
Licences				Add Licence		
Licensing Summary						
Source Licensing				Feature Licensing		
Source Type ^	Status ^	Tagging Write Back ^	Valid To ^	Feature Name ^	Status ^	Valid To ^
Box	✓ Valid	⌵	2019-04-04	Custom Reports	✓ Valid	2021-01-15
CMIS	✓ Valid	⌵	2019-04-04	Redaction	✓ Valid	2019-04-04
Content Server	✓ Valid	⌵	2019-04-04	Reporting	✓ Valid	2021-01-15
Exchange	✓ Valid	⌵	2019-04-04	Workflows	✓ Valid	2019-04-04
File	✓ Valid	⌵	2021-01-15			
Google Drive	✓ Valid	⌵	2019-04-04			
Outlook Archive	✓ Valid	⌵	2019-04-04			
Salesforce	✓ Valid	⌵	2021-01-15			
SharePoint	✓ Valid	⌵	2021-01-15			
SQL	✓ Valid	⌵	2019-04-04			
Web	✓ Valid	⌵	2019-04-04			

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:

Delete		
<input type="checkbox"/> Field	Ignore	Search...
<input type="checkbox"/> _IsCurrentVersion	<input type="checkbox"/>	Edit Delete
<input type="checkbox"/> _ModerationStatus	<input type="checkbox"/>	Edit Delete
<input type="checkbox"/> _reportinggallerymetadataid	<input type="checkbox"/>	Edit Delete
<input type="checkbox"/> _reportinggallerytemplateid	<input type="checkbox"/>	Edit Delete
<input type="checkbox"/> Abstract	<input type="checkbox"/>	Edit Delete
<input type="checkbox"/> allowslistpolicy	<input type="checkbox"/>	Edit Delete
<input type="checkbox"/> AnonymousViewMask	<input type="checkbox"/>	Edit Delete
<input type="checkbox"/> Attachments	<input type="checkbox"/>	Edit Delete
<input type="checkbox"/> BaseType	<input type="checkbox"/>	Edit Delete

Metadata Field Mappings

This table allows additional metadata fields to be generated by mapping an already existing field name to a new name.

Delete		Add
<input type="checkbox"/> Source Field Name ▲	Target Field Name ▼	Search...
<input type="checkbox"/> Title	PageTitle	Edit Delete
Copy CSV XLSX		Showing 1 record(s) Page Size: 10 25 50 100 200

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.

Metadata Value Mappings			
Delete		Add	
<input type="checkbox"/> Field Name(s) ▲	Source Value ↕	Target Value ↕	Search...
<input type="checkbox"/> Title	Test	Redacted	Edit Delete
<div> Copy CSV XLSX </div> <div>Showing 1 record(s)</div> <div>Page Size: 10 25 50 100 200</div>			

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**.

Email Servers					
 Delete		Add			
<input type="checkbox"/>	Host 	Port 	Use SSL 	From Email 	Username 
<input type="checkbox"/>	smtp.office365.com	587		noreply@conceptsearching.com	demo@conceptsearching.com
Edit Delete					
 Copy CSV XLSX					
Showing 1 record(s)					
Page Size: 10 25 50 100 200					

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 Server Details

Email Server Details

Host:

smtp.office365.com

Port:

587

☒ Use SSL

From Email:

demo@conceptsearching.com

Username:

demo@conceptsearching.com

Password:

.....

Test Configuration Settings

We recommend you test your configuration by entering a confirmation email address below.

Email Address:

test@conceptsearching.com

Save

Cancel

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.

Email Group Details

Group Name:

Test

Email Addresses:

+

✕ demo@conceptsearching.com

✕ admin@conceptsearching.com

Email Server:

smtp.office365.com:587 (mikep@conceptsearching.com)

Save

Cancel

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.

Health Notifications			
<div> Delete </div>			
<input type="checkbox"/>	Group Name ▲	Notification Mode ▼	Daily Summary ▼
<input type="checkbox"/>	Test	Warning And Error	<input checked="" type="checkbox"/>
<div> Edit Delete </div>			
<div> Copy CSV XLSX </div>			
<div> Showing 1 record(s) </div>			
<div> Page Size: 10 25 50 100 200 </div>			

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.

Health Notification Configuration

Group Name:

Test

Notification Mode:

☐ Off
☐ Errors
☒ Errors & Warnings

Daily Summary:

☒

Save

Cancel

7.2.5. Text Handling

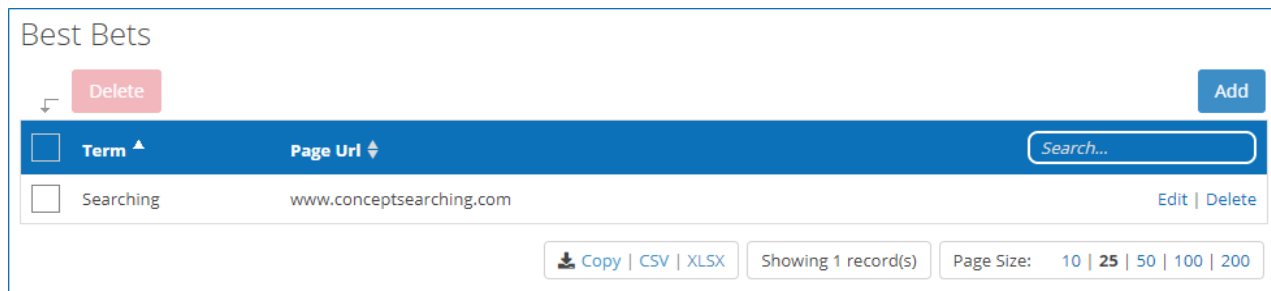
This section contains information on how to configure text processing. Review the following for additional information:

- [Best Bets](#)
- [Content Type Extension Mapping](#)
- [Content Type Extraction Methods](#)
- [Language Detection](#)
- [No Stem](#)
- [OCR Language Mapping](#)

- [Synonyms](#)
- [Text Patterns](#)

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.



Best Bets

Delete Add

Term	Page Url	
Searching	www.conceptsearching.com	Edit Delete

Copy | CSV | XLSX Showing 1 record(s) Page Size: 10 | 25 | 50 | 100 | 200

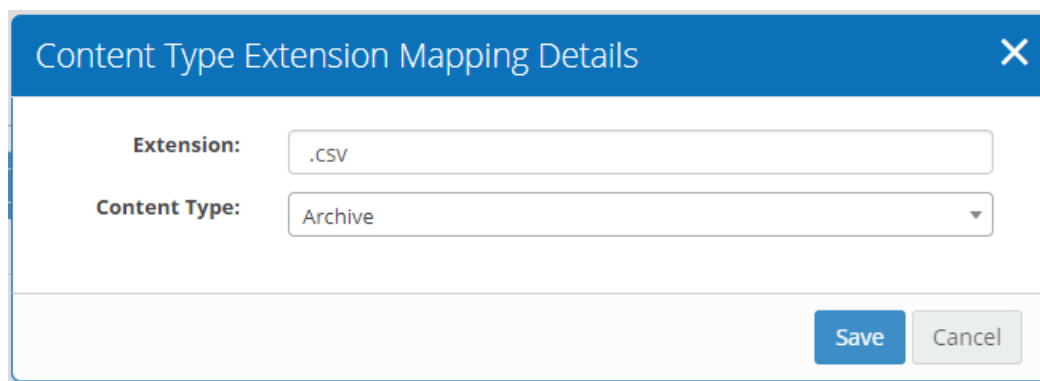
First, enter the search term that you wish to match and then click the **Add** button.

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 Extension Mapping Details

Extension: .CSV

Content Type: Archive

Save Cancel

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.

Content Type Extraction Methods				
iFilters: Click here to view the list of installed iFilters.				
Content Type ^	Default Extension ^	Extraction Method ^	Use IFilter as Backup ^	Search...
Adobe Photoshop	.psd	iFilter	<input type="checkbox"/>	Re-Index Edit
AIFF	.aiff	iFilter	<input type="checkbox"/>	Re-Index Edit
Archive	.zip	iFilter	<input type="checkbox"/>	Re-Index Edit
Bitmap	.bmp	iFilter	<input type="checkbox"/>	Re-Index Edit
CAD	.dwg	Aspose	<input type="checkbox"/>	Re-Index Edit
Compiled HTML	.chm	iFilter	<input type="checkbox"/>	Re-Index Edit
DICOM	.dcm	iFilter	<input type="checkbox"/>	Re-Index Edit

Language Detection

The language detection list specifies which languages will be considered for auto-detection.

Language Detection

☒ All Languages

<input checked="" type="checkbox"/> Afrikaans (af)	<input checked="" type="checkbox"/> Albanian (sq)	<input checked="" type="checkbox"/> Arabic (ar)	<input checked="" type="checkbox"/> Aragonese (an)	<input checked="" type="checkbox"/> Asturian (ast)
<input checked="" type="checkbox"/> Basque (eu)	<input checked="" type="checkbox"/> Belarusian (be)	<input checked="" type="checkbox"/> Bengali (Bangla) (bn)	<input checked="" type="checkbox"/> Breton (br)	<input checked="" type="checkbox"/> Bulgarian (bg)
<input checked="" type="checkbox"/> Catalan (ca)	<input checked="" type="checkbox"/> Chinese (Simplified) (zh-cn)	<input checked="" type="checkbox"/> Chinese (Traditional) (zh-tw)	<input checked="" type="checkbox"/> Croatian (hr)	<input checked="" type="checkbox"/> Czech (cs)
<input checked="" type="checkbox"/> Danish (da)	<input checked="" type="checkbox"/> Dari (prs)	<input checked="" type="checkbox"/> Dutch (nl)	<input checked="" type="checkbox"/> English (en)	<input checked="" type="checkbox"/> Estonian (et)
<input checked="" type="checkbox"/> Finnish (fi)	<input checked="" type="checkbox"/> French (fr)	<input checked="" type="checkbox"/> Galician (gl)	<input checked="" type="checkbox"/> German (de)	<input checked="" type="checkbox"/> Greek (el)
<input checked="" type="checkbox"/> Gujarati (gu)	<input checked="" type="checkbox"/> Haitian (ht)	<input checked="" type="checkbox"/> Hebrew (he)	<input checked="" type="checkbox"/> Hindi (hi)	<input checked="" type="checkbox"/> Hungarian (hu)
<input checked="" type="checkbox"/> Icelandic (is)	<input checked="" type="checkbox"/> Indonesian (id)	<input checked="" type="checkbox"/> Irish (ga)	<input checked="" type="checkbox"/> Italian (it)	<input checked="" type="checkbox"/> Japanese (ja)
<input checked="" type="checkbox"/> Kannada (kn)	<input checked="" type="checkbox"/> Khmer (km)	<input checked="" type="checkbox"/> Korean (ko)	<input checked="" type="checkbox"/> Latvian (lv)	<input checked="" type="checkbox"/> Lithuanian (lt)
<input checked="" type="checkbox"/> Macedonian (mk)	<input checked="" type="checkbox"/> Malay (ms)	<input checked="" type="checkbox"/> Malayalam (ml)	<input checked="" type="checkbox"/> Maltese (mt)	<input checked="" type="checkbox"/> Maori (mi)
<input checked="" type="checkbox"/> Marathi (mr)	<input checked="" type="checkbox"/> Nepali (ne)	<input checked="" type="checkbox"/> Norwegian (no)	<input checked="" type="checkbox"/> Occitan (oc)	<input checked="" type="checkbox"/> Panjabi (Punjabi) (pa)
<input checked="" type="checkbox"/> Pashto (ps)	<input checked="" type="checkbox"/> Persian (Farsi) (fa)	<input checked="" type="checkbox"/> Polish (pl)	<input checked="" type="checkbox"/> Portuguese (pt)	<input checked="" type="checkbox"/> Romanian (ro)
<input checked="" type="checkbox"/> Russian (ru)	<input checked="" type="checkbox"/> Serbian (sr)	<input checked="" type="checkbox"/> Slovak (sk)	<input checked="" type="checkbox"/> Slovene (sl)	<input checked="" type="checkbox"/> Somali (so)
<input checked="" type="checkbox"/> Spanish (es)	<input checked="" type="checkbox"/> Swahili (sw)	<input checked="" type="checkbox"/> Swedish (sv)	<input checked="" type="checkbox"/> Tagalog (tl)	<input checked="" type="checkbox"/> Tamil (ta)
<input checked="" type="checkbox"/> Telugu (te)	<input checked="" type="checkbox"/> Thai (th)	<input checked="" type="checkbox"/> Turkish (tr)	<input checked="" type="checkbox"/> Ukrainian (uk)	<input checked="" type="checkbox"/> Urdu (ur)
<input checked="" type="checkbox"/> Vietnamese (vi)	<input checked="" type="checkbox"/> Welsh (cy)	<input checked="" type="checkbox"/> Yiddish (yi)		

Save

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

Delete

Add

<input type="checkbox"/> Term ▲	
<input type="checkbox"/> CO	Edit Delete
<input type="checkbox"/> cos	Edit Delete
<input type="checkbox"/> finish	Edit Delete
<input type="checkbox"/> Finnish	Edit Delete
<input type="checkbox"/> GA	Edit Delete
<input type="checkbox"/> gas	Edit Delete
<input type="checkbox"/> international	Edit Delete

Showing 7 record(s)

Page Size: 10 | 25 | 50 | 100 | 200

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

Delete

Add

<input type="checkbox"/> Inclusion Filter ▲	Mapping ▼	
<input type="checkbox"/> */fr/*	French (fr)	Edit Delete

Showing 1 record(s)

Page Size: 10 | 25 | 50 | 100 | 200

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

Delete

Add

<input type="checkbox"/> Term ▲	Synonyms ▼	
<input type="checkbox"/> Searching	Concept Searching	Edit Delete

Copy

CSV

XLSX

Showing 1 record(s)

Page Size: 10 | 25 | 50 | 100 | 200

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.

Text Patterns				
Delete				Add
<input type="checkbox"/>	Start Tag ^	End Tag ^	Tag Type ^	Doc Types ^
<input type="checkbox"/>	E.ON		INDEX TERM	Text
				Edit Delete
<div> Copy CSV XLSX Showing 1 record(s) Page Size: 10 25 50 100 200 </div>				

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				
Delete		Add		
<input type="checkbox"/>	Plan Name ^	NLP Redaction ▾	Regex Redaction ▾	Redaction Groups ▾
<input type="checkbox"/>	Complete	▼	▼	Columns
<input type="checkbox"/>	Test Regex	■	▼	
Copy CSV XLSX Showing 2 record(s) Page Size: 10 25 50 100 200				

Redaction Entities

Redaction entities can be used to specify any custom words or phrases that should be removed by a redaction plan.

Entities		
Delete		Add
<input type="checkbox"/>	Entity Name ^	Entity Group ▾
<input type="checkbox"/>	Retention Schedule	Columns
<input type="checkbox"/>	Sensitive	None
Copy CSV XLSX Showing 2 record(s) Page Size: 10 25 50 100 200		

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

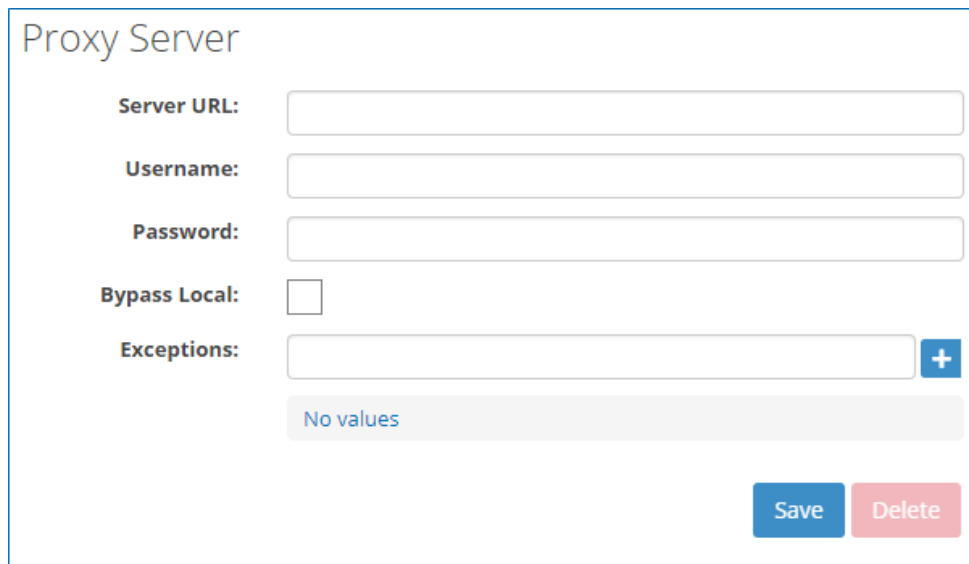
Sometimes an application may wish to remove selected documents from all search results. This may be implemented by specifying **No Index** entries.

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.

The screenshot shows a web form titled "Proxy Server". It contains five input fields: "Server URL:", "Username:", "Password:", "Bypass Local:" (a checkbox), and "Exceptions:" (a text input with a blue "+" button to its right). Below the "Exceptions:" field is a light gray box containing the text "No values". At the bottom right of the form are two buttons: a blue "Save" button and a red "Delete" button.

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.

Suspend Service Details [X]

Source: All Sources

Service: All Services

Day: Any Day

Start Time: 01:30 [Clock Icon] **End Time:** 02:00 [Clock Icon]

[Save] [Cancel]

01 : 30

00 13 14 15 16 17 18 19 20 21 22 23

12 11 10 9 8 7 6 5 4 3 2

Set

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/Times**—Allows the configuration of which days and times the suspension will be in place.

7.2.8. Configuration Backup

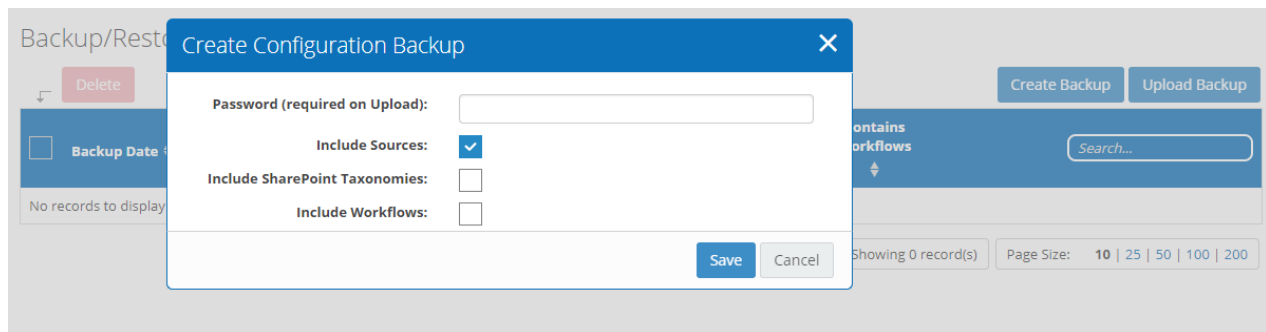
NOTE: Only available for 'Superusers'

The **Backup** utility allows for the migration of complex Netwrix Data Classification instance configurations.

This allows a user to safely design and test a conceptSearching 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)

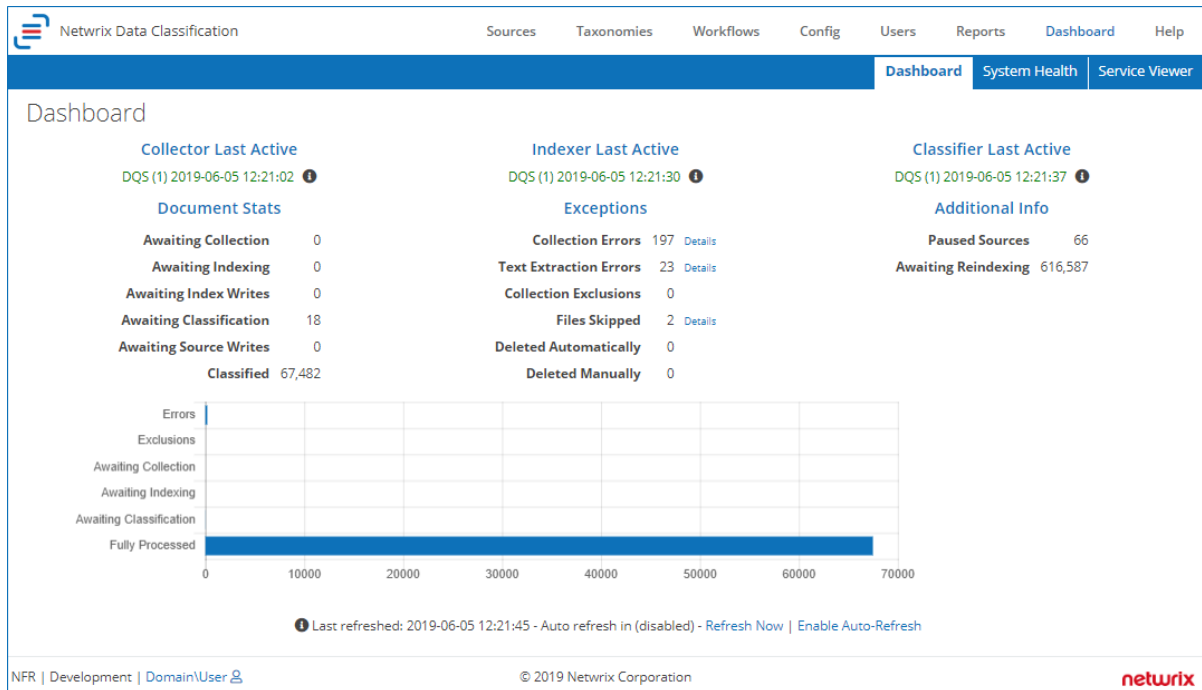
Statistics shown on the **Dashboard** screen are cached and updated regularly by the Collector service. If the values are not being updated please ensure that the Collector service is running.

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

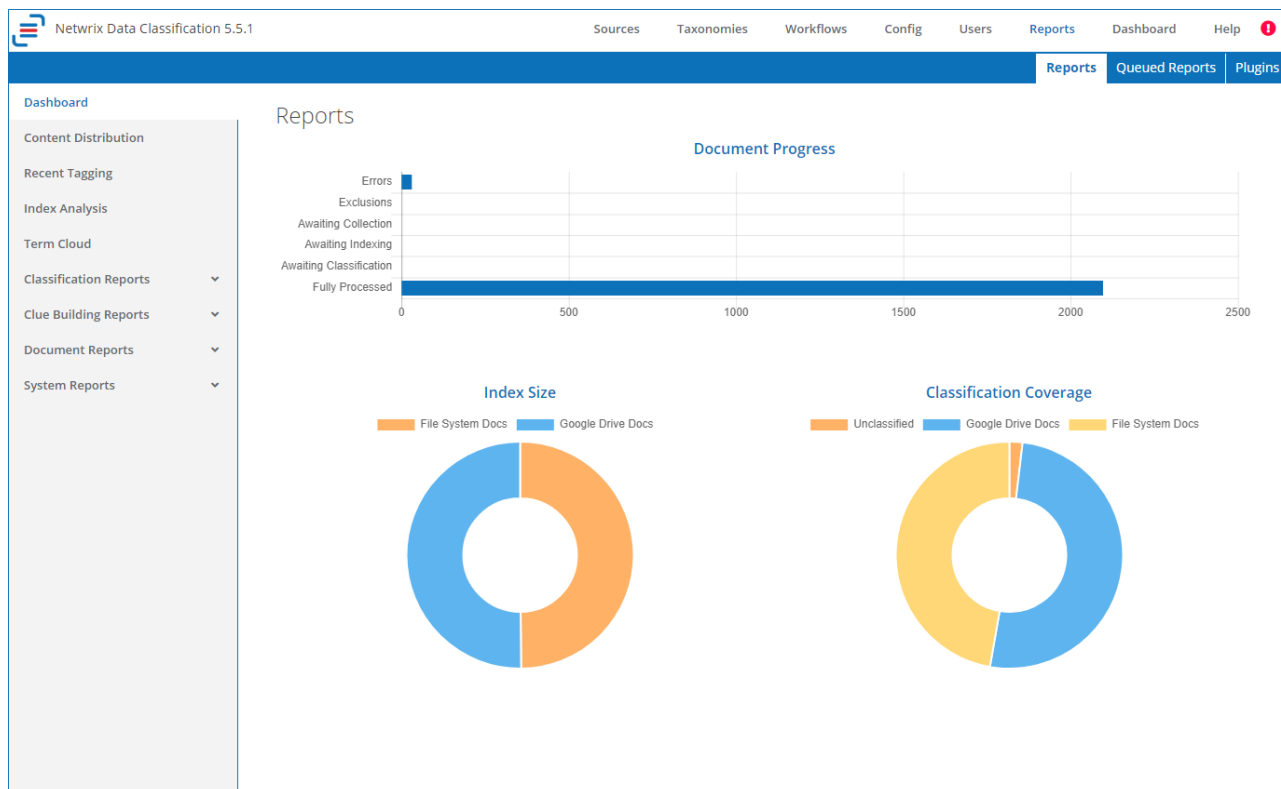
From the **Netwrix Data Classification Service Viewer** it is possible to view a live stream of the current work being processed by the Windows Services. As the services progress each document the display will change. Once all work is complete "Idle..." will be displayed.

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 Netrix 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:

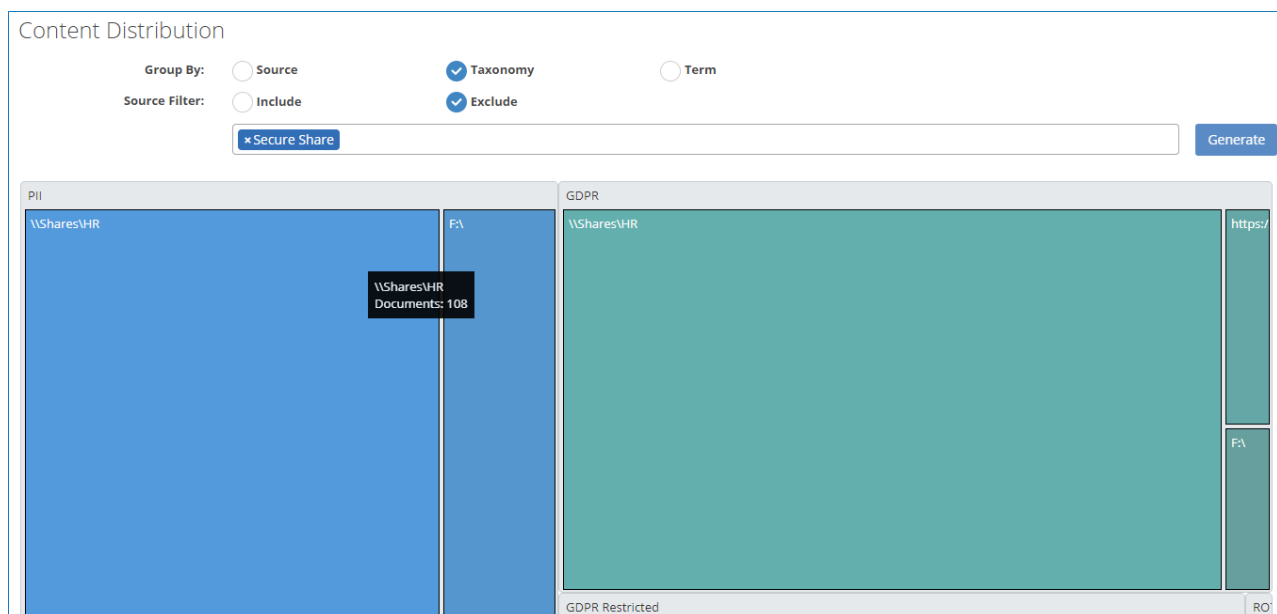
- [Content Distribution](#)
- [Review Built-in Reports](#)
- [Types of Reports](#)

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

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 Indexer Settings and rebuild the desired sources.

Match Precision (%): 95
Minimum Text Length: 100
Maximum Text Length Difference: 20

- Hide filters

Page URL:
Source:
Restrict the report to one or more source
Exclude Content Types:
Removes documents matching the specified content types from the report results (comma delimited)

PageId	PageUrl	Last Modified	Collect Date	Status	Duplicate PageId	Duplicate PageUrl	Relevancy	Text Length Difference	Generated At
51	https://2010.conceptsearching.com/Shared Documents/02975.pdf	2016-04-20 14:05:02	2019-02-27 15:35:39	400	59	https://2010.conceptsearching.com/Shared Documents/Test/02975.pdf	100	0.00	2019-03-05 11:52:29
53	https://2010.conceptsearching.com/Shared Documents/BandB Press Release.DOC	2016-04-20 14:05:10	2019-02-27 15:35:40	400	61	https://2010.conceptsearching.com/Shared Documents/Test/BandB Press Release.DOC	100	0.00	2019-03-05 11:52:30
54	https://2010.conceptsearching.com/Shared Documents/CC doc2.docx	2016-04-20 14:05:13	2019-02-27 15:35:39	400	63	https://2010.conceptsearching.com/Shared Documents/Test/CC doc2.docx	100	0.00	2019-03-05 11:52:30
55	https://2010.conceptsearching.com/Shared Documents/Integrity Matters.docx	2016-04-20 14:05:15	2019-02-27 15:35:40	400	74	https://2010.conceptsearching.com/SubSite/Shared Documents/Integrity Matters.docx	100	0.00	2019-03-05 11:52:30
55	https://2010.conceptsearching.com/Shared Documents/Integrity Matters.docx	2016-04-20 14:05:15	2019-02-27 15:35:40	400	64	https://2010.conceptsearching.com/Shared Documents/Test/Integrity Matters.docx	100	0.00	2019-03-05 11:52:30
56	https://2010.conceptsearching.com/Shared Documents/newsletter 0518.pdf	2016-04-20 14:05:16	2019-02-27 15:35:40	400	65	https://2010.conceptsearching.com/Shared Documents/Test/newsletter 0518.pdf	100	0.00	2019-03-05 11:52:30

Certain document specific reports can be exported along with any associated document metadata.

Export

The selected report can be exported along with any associated document metadata. Simply enter the metadata values you wish to export below (multi-value fields will be delimited with a semicolon)

Include Metadata? ☒

Metadata Fields:

To export specific metadata:

1. First choose one the 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

Provides a list of documents tagged with a given set of classifications. For full debugging detail the trace mode 'Classification Calculations' must be enabled prior to auto-classification.

Classification:

Mode:

Include detail? ☒

- Hide filters

Page URL:

Source:

Restrict the report to one or more source

Generate

PageId ^	PageUrl ↕	Page Title	Term Name	Score	Boost Score	Clue	Clue Type	Clue Score	Info
40	https://2010.conceptsearching.com/Shared Documents/1044_venezuela_and_cover_changes.pdf		Environment	70	2	Natural resources	Standard	46	Count=2
						Environment	Standard	13	Count=7
						conservation	Standard	11	Count=2

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.

Queued Reports

[Delete](#)

<input type="checkbox"/>	Request Date	Title	Type	Status	Requestor	Processed Date	Time (secs)	Search...
<input type="checkbox"/>	2019-03-05 12:19:05	Health, well-being and care (IPSV) - "Defence medical services" Health "social care" "Health and Wel	Document Search Results Extract	Pending	DEMO			Delete
<input type="checkbox"/>	2019-03-05 12:18:36	Environment (IPSV) - conservation protection Wildlife "Natural resources" "environmental protection"	Document Search Results Extract	Pending	DEMO			Delete

[Copy](#) | [CSV](#) | [XLSX](#) Showing 2 record(s) Page Size: **10** | 25 | 50 | 100 | 200

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.

Plugins

Plugins will be detected in the following location: C:\Program Files\ConceptSearching\Plugins\

[Detect Plugins](#)

Name	Assembly Path	Error	Enabled	Search...
ClassificationReport	C:\Program Files\ConceptSearching\Plugins\conceptReportsClassification.dll		<input checked="" type="checkbox"/>	Edit
PageStatusReport	C:\Program Files\ConceptSearching\Plugins\conceptReportsPageStatus.dll		<input checked="" type="checkbox"/>	Edit

[Copy](#) | [CSV](#) | [XLSX](#) Showing 2 record(s) Page Size: **10** | 25 | 50 | 100 | 200