Preservica

System Administration Guide

v6.6.1

Table of Contents

References

Document	Ref	Date	Details & Issue
Preservica Guide to System Documentation	[DOC]	Sept 2022	svn/doc/UD/GSD V5.R9.M0
System Installation Guide	[SIG]	See [DOC] for version information.	
Preservica System User Guide	[SUG]		
Preservica System Maintenance guide	[SMG]		
Technical Description	[TD]		
Preservica Information Package Structure Definition	[SIP]		
Preservica Standard Workflows	[SWF]		
Preservica Developer Guide	[DEV]		
Preservica Storage Adapters	[STORE]		
SIP Creator System User Guide	[SCUG]		

Chapter 1. Introduction

1.1. Purpose of this document

This document describes how to use the administration and management functions of the Preservica Digital Preservation System. It is intended for use by system administration or management staff only. This document does not cover the installation and initial configuration of the Preservica system or system maintenance activities which are covered in separate documents.

1.2. Scope of this document

This document covers the core administrative and management functionality of the Preservica system that is available within all system installations. Extensions to the core system for specific customers will be covered in a separate customer specific document or incorporated into a customer specific version of this document.

1.3. Context of this Issue

This version of the SAG corresponds to the Preservica 6.6 release.

1.4. Definition of Terms

Term	Definition
SUG	The System User Guide, describes how to use the standard features of the system.
SAG	System Administration Guide describes how to use the administration features of the system
SIG	The System Installation Guide, describes how to install the system and initial system configuration.
SMG	The System Maintenance Guide, describes how to maintain the system.
Preservica	Preservica Enterprise edition
LDR	Linked Data Registry
PUID	Persistent Unique IDentifier (allocated in PRONOM)
GUID	Globally Unique IDentifier
BPMN	Business Process Modelling Notation
CSV	Comma Separated Values
FTP	File Transfer Protocol, a standard means of transferring files from a remote server
SIP	Submission Information Package
DIP	Dissemination Information Package
wsdl	Web Services Description Language. An XML format used to describe web services and how to access them (by defining the 'contract' between the service and the user).

Chapter 2. Overview of the Preservica System

The Preservica digital preservation system has been built specifically to deal with the problems of digital preservation in libraries, archives and other organisations facing the problems of long-term retention of digital content (sufficiently long-term that the material will inevitably suffer from the obsolescence of the software or hardware used to create it).

Its primary purpose is to retain born-digital or digitised information objects permanently and provide access to these information objects in current technologies.

Preservica is based on the Open Archival Information System (OAIS) reference model. The functional entities of this are summarised below:

- **Ingest**: These are the steps required to transfer items from their current location into the archive in a managed manner.
- **Archival Storage**: The storage of the bulk data (usually files) based on off-the-shelf storage management tools.
- **Data Management**: Tools to manage the storage of the archive, including the metadata.
- **Administration**: A set of tools to administer the system and access to it.
- **Access**: Tools to search, browse and download the contents of the archive.
- **Preservation**: The unique module that manages the information so that it can be accessed long into the future.

The Preservica system is a collection of, mainly web-based, applications. The Preservica system takes collections of files as its primary input and then processes these files to extract technical file information (metadata) to the database, store these files securely within the file storage system, and provide the user with an interface to interact with the stored files to ensure their long term accessibility, preserving them in perpetuity.

The Preservica System comprises the following components, some of which are documented separately:

- **Workflow Engine -** The workflow application that contains the user interfaces and business logic of the Preservica system. Individual workflows can be configured as required to meet specific business needs where necessary through the creation of additional workflow steps.
- **Job Queue** - JobQueue is the Preservica work engine. To keep the Preservica user interface responsive, the Preservica Workflow Engine does not perform any long running jobs itself; instead it passes these over to JobQueue for processing, and waits for JobQueue to notify the Preservica Workflow Engine once the job is complete.
- **Explorer** - The Preservica Browser application allows the user to browse the hierarchy of folders, and view assets, stored in the archive and view the metadata associated with them. Many workflows and other operations can also be started from Explorer.
- **Search** - Preservica has a search interface that allows the user to search the metadata associated with collections, deliverable units and files, together with any full text extracted from the archived files.
- **Registry** - The technical registry provides access to information about tools and services that support preservation risk assessment, migration pathway planning, object identification and validation and metadata extraction.
- **Database** - Preservica stores archiving metadata, audit information and run time information in a relational database. All access to the database is via the Preservica applications and invisible to system users.
- **File Store -** Preservica stores the digital files in a file store. This may be a simple directory structure or a proprietary bulk storage system, e.g. Amazon S3.

- **Upload and Preparation Tool** (PUT) - A web application that allows end users to create SIPs without the need for standalone client side tools.

- **SIP Creator** and **Upload Wizard** - Standalone Java desktop applications that can be used by end users to create a Submission Information Package (SIP) ready for ingest by Preservica. It will create the XML metadata file that contains the structural metadata describing the files and directories within a SIP.

- **Universal Access**, a Wordpress-based public portal to provide a simplified, read-only view of the archive, to Preservica users or the general public.

- **Authentication Service** - An external component to the Preservica system; typically it is maintained by the client's IT department. Preservica communicates with the Authentication service to validate user login credentials (typically username and password). Once authenticated, the user's permissions to access the Preservica functionality based on their association to groups within the **Authentication Service** is provided back to Preservica.

Chapter 3. Workflows

A key element of the Preservica system is the highly flexible and configurable workflow engine. To aid understanding of key areas of functionality within Preservica that make use of the workflow engine, such as ingest, preservation and access, it is necessary to define some key terms first.

A **Workflow Definition** is a specific arrangement of a number of (small) processing tasks (or **workflow steps**) to achieve a specific business need. Preservica is supplied with a number of standard workflow steps and standard workflow definitions, which your system administrator should upload.

Typically, workflow definitions will be loaded when Preservica is installed and will be changed infrequently thereafter. Workflow definitions are created in the form of XML files, and the ability to upload these files into the Preservica system is restricted to the Preservica Administration role (`SDB_ADMIN_USER`).

Each workflow definition has a defined **Workflow Type**, which indicates the functional area of the system that it applies to (i.e. ingest, preservation, access, data management).

The term **Workflow Context** is used to describe the combination of a workflow definition and the configuration information required to enable it to be run. For an ingest workflow, the workflow context would typically define the location of the files to process, operator (user) notification details and possibly scheduling information. Typically several workflow contexts may be created based on the same workflow definition with each workflow context meeting specific processing requirements. Continuing the ingest workflow example, if the material to be ingested is supplied in several different locations, then several workflow contexts could be created from the same workflow definition to handle this: one for each source location.

A **Workflow Instance** is a specific, single execution of a workflow context, regardless of whether it was started manually by a user or automatically by the Preservica system (in the case of scheduled workflows). Actual system processing (ingest or preservation actions) is undertaken by workflow instances. Each workflow instance stores the time that it started, the time it finished, its state and the progress of the step instances associated with it.

Each workflow instance consists of one or more **Step Instances**: specific executions of individual workflow steps. Each step instance stores the time the step started, the time it finished, its state, and any errors that occurred during its execution.

3.1. Workflow Definitions

Workflow definitions define the steps within a workflow and their order; typically only a small number of workflow definitions will be used within a single Preservica system.

The **Workflow Definitions** tab shows a list of registered workflows (organised by OAIS functional area) and allows additional workflow definitions to be uploaded into Preservica. For each workflow definition its name, version, and the date when it was uploaded and by whom are shown. The **Download** button allows you to download a copy of a workflow definition.

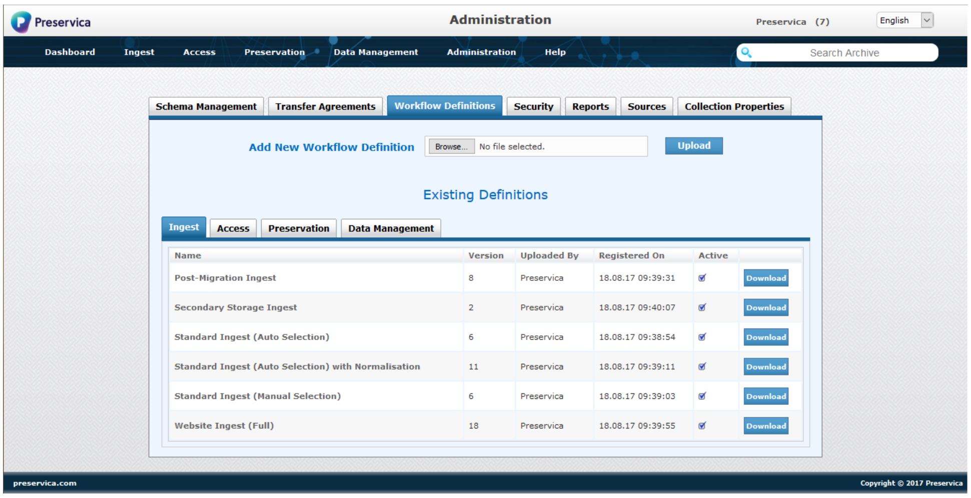
In addition, the **Active** checkbox for each workflow definition gives the administrator control over which workflow definitions can be used in the system. Managers can create workflow contexts from workflow definitions *only* if the workflow definition is active. By default, when a workflow definition is uploaded, it is activated. If a workflow definition is inactive, then it will not appear in the list of available workflow definitions in the manage tab of the appropriate OAIS functional area. If an administrator deactivates a workflow definition (by unticking the active checkbox), then any workflow contexts created from it will be deactivated and it is not possible to reactivate them from the appropriate manage tab and users will receive a warning message if they try to do so. If a workflow definition is reactivated, then any workflow contexts created from it will remain deactivated, until activated individually by a user with appropriate access.

note

Adding new workflow definitions is only available to administrators, not managers.

To add a new workflow definition, press the **Browse…** button, which will open a browse window to allow you to locate and select the file to upload. The Workflow definition (.rf) file must be located on a drive accessible to your web browser. Once you have selected your workflow definition, press the **Upload** button to upload it to Preservica. Preservica will validate the syntax of the workflow definition and reject any invalid files. Preservica will also check that the workflow definition is unique, that is the combination of its ID and its version is unique within the system, and reject any workflow definitions that have been uploaded previously.

Figure 3.1. Workflow Definition List



Add New Workflow Definition No file selected.

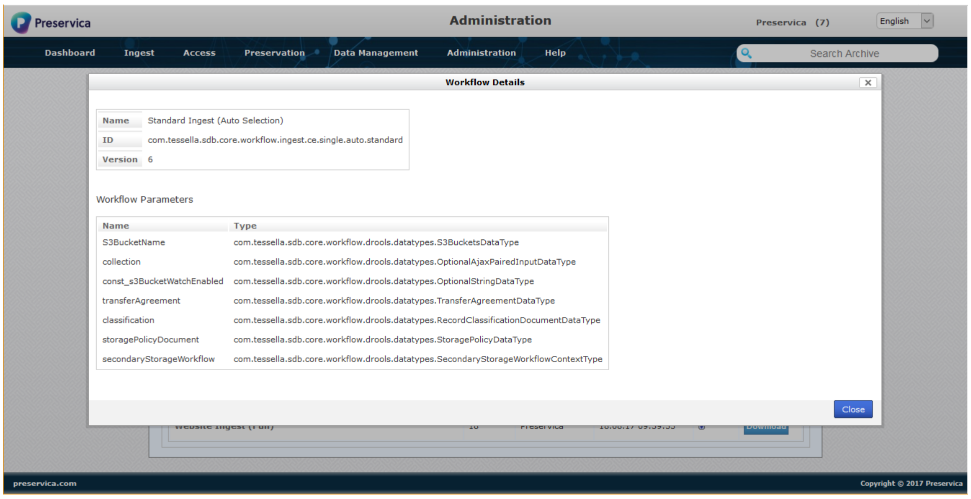
Existing Definitions

Name	Version	Uploaded By	Registered On	Active	
Post-Migration Ingest	8	Preservica	18.08.17 09:39:31	☑	Download
Secondary Storage Ingest	2	Preservica	18.08.17 09:40:07	☑	Download
Standard Ingest (Auto Selection)	6	Preservica	18.08.17 09:38:54	☑	Download
Standard Ingest (Auto Selection) with Normalisation	11	Preservica	18.08.17 09:39:11	☑	Download
Standard Ingest (Manual Selection)	6	Preservica	18.08.17 09:39:03	☑	Download
Website Ingest (Full)	18	Preservica	18.08.17 09:39:55	☑	Download

Clicking on the **workflow definition** name (as displayed in the **Name** column) will display summary information (the name, ID, version and parameters) for the workflow definition (see Figure 3.2); this can be used to confirm the correct file has been uploaded.

It is possible to upload a new version of an existing workflow definition. The new Workflow definition (RF) file must explicitly specify a version number that is higher than the version currently loaded into the Preservica system (see [DEV] for more details). When the new version is loaded it replaces the existing version. If any active workflow contexts exist based on the workflow definition they will be deactivated and a user (as specified in the workflow context) notified by e-mail to ensure that the workflow context is reviewed before use.

Figure 3.2. Workflow Definition Details



Name	Standard Ingest (Auto Selection)
ID	com.tessella.sdb.core.workflow.ingest.ce.single.auto.standard
Version	6

Workflow Parameters

Name	Type
S3BucketName	com.tessella.sdb.core.workflow.drools.datatypes.S3BucketsDataType
collection	com.tessella.sdb.core.workflow.drools.datatypes.OptionalAjaxPairedInputDataType
const_s3BucketWatchEnabled	com.tessella.sdb.core.workflow.drools.datatypes.OptionalStringDataType
transferAgreement	com.tessella.sdb.core.workflow.drools.datatypes.TransferAgreementDataType
classification	com.tessella.sdb.core.workflow.drools.datatypes.RecordClassificationDocumentDataType
storagePolicyDocument	com.tessella.sdb.core.workflow.drools.datatypes.StoragePolicyDataType
secondaryStorageWorkflow	com.tessella.sdb.core.workflow.drools.datatypes.SecondaryStorageWorkflowContextType

Workflow Contexts customise the workflow for specific requirements. Typically multiple **Workflow Contexts** based on the same definition will be used. Workflow contexts are created from the **Manage** tab on the main **Ingest**, **Preservation**, **Access**, or **Data Management** screen.

3.2. Workflow Management

Workflow functionality is all accessed from the main workflow management screens, which is reached by selecting the relevant functional area (Ingest, Access, Preservation or Data Management) by its menu item in the header. Each tab is also accessible from the menu which appears when hovering over or tabbing onto the menu items. To access it, you need to have a suitable system role: either `SDB_MANAGER_USER` or `SDB_ADMIN_USER`, or the relevant functional role for the section (e.g. `SDB_INGEST_USER` to access the Ingest workflows).

On the workflow management pages there are several tabs:

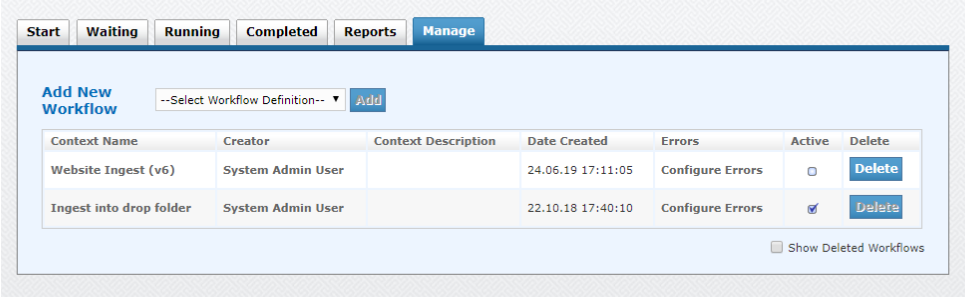
- The **Manage** tab will allow you to create and edit*workflow contexts*. In order to edit an existing context, you must first deactivate it. Before being able to run a context (from the Start tab, Explorer or via an API call), you must activate it.
- The **Start** tab will allow you to start a specific *workflow instance* manually. Some workflow definitions will create contexts that can't be started manually, either because they are designed to be run from Explorer on particular entities, or because they are started internally. See the documentation for each workflow in the Standard Workflows document for more details.
- The **Waiting** tab will show you a list of **workflow instances** that are waiting for user action: typically either the selection of a SIP to ingest or a decision on what to do about an encountered exception is required.
- The **Running** tab will show you a list of **workflow instances** that are running and allow you to monitor their progress.
- The **Completed** tab will show you a list of **workflow instances** that have finished and allow you to review the process and any related log files.
- The **Reports** tab will allow you to run any ingest-related reports that have been loaded into the system.

3.2.1. Creating Workflow Contexts

The workflow definition defines the steps in the workflow and their order; typically only a few workflow definitions will be used within a single Preservica system. **Workflow Contexts** customise a workflow for specific requirements, e.g. by specifying the folder from which SIPs (Submission Information Packages) will be ingested. Several **Workflow Contexts** based on the same definition may be used within a single Preservica system, with different parameters.

Workflow contexts are created from the **Manage** tab on the main workflow management screen (see Figure 3.3). To create a new workflow context, select a workflow definition from the drop down list and then press the **Add** button.

Figure 3.3. The manage tab of the main ingest screen

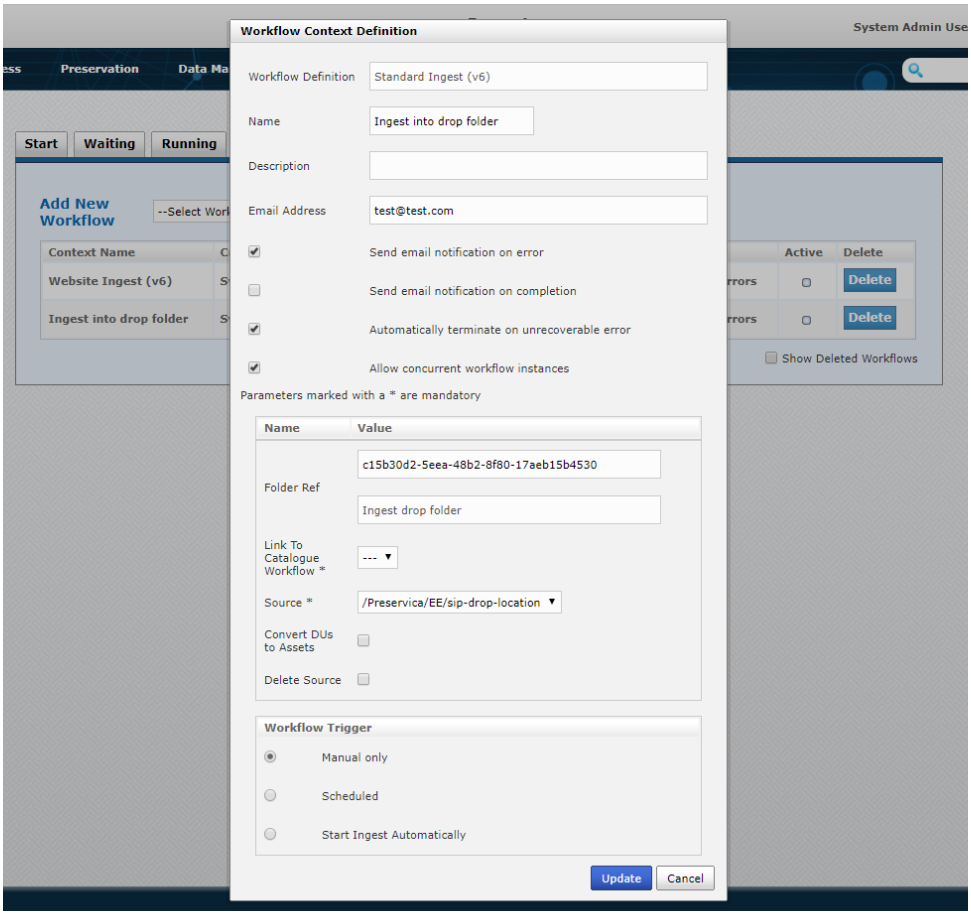


Context Name	Creator	Context Description	Date Created	Errors	Active	Delete
Website Ingest (v6)	System Admin User		24.06.19 17:11:05	Configure Errors	□	Delete
Ingest into drop folder	System Admin User		22.10.18 17:40:10	Configure Errors	☑	Delete

□ Show Deleted Workflows

The system will then display the **Define Workflow Context** dialog box (see Figure 3.4) into which you can enter the necessary information to define the workflow context. The information required will depend on the workflow definition so the appearance of the dialog box will be different for different workflows.

Figure 3.4. Ingest workflow context definition dialog box



Workflow Context Definition

Workflow Definition: Standard Ingest (v6)

Name: Ingest into drop folder

Description:

Email Address: test@test.com

☑ Send email notification on error

☐ Send email notification on completion

☑ Automatically terminate on unrecoverable error

☑ Allow concurrent workflow instances

Parameters marked with a * are mandatory

Name	Value
Folder Ref	c15b30d2-5eea-48b2-8f80-17aeb15b4530 Ingest drop folder
Link To Catalogue Workflow *	---
Source *	/Preservica/EE/sip-drop-location
Convert DUs to Assets	☐
Delete Source	☐

Workflow Trigger

● Manual only

○ Scheduled

○ Start Ingest Automatically

Specific workflow definitions require specific sets of information to be defined in the workflow context. The Preservica Standard Workflows [SWF] document describes what information is required for each of the standard workflows supplied with Preservica. However there are some common parameters available for all workflow contexts:

- A **Name** for the workflow context. This is displayed (as the context name) on the Running and other tab to identify the workflow context from which an instance was created. It is also the name that will appear in the Actions context menu, the Actions menu in Search and on the Actions tab in the Explorer properties view, if the workflow definition in question can be run from there.

- A more detailed, free-text **Description** of the workflow context.

- An **Email Address** of a user (or operator) to inform of the progress of any **workflow instances** based on this context. This should not be confused with email notifications to an information supplier (publisher or government department) that a SIP has been ingested successfully. These email notifications form part of the business process and are included in the workflow definition as a specific notification step.

- Email notification indicators: these allow the user to indicate whether the operator should be notified in certain circumstances, such as the completion of the workflow or if an error is encountered, for any **workflow instances** based on this context.

- Automatic termination indicator: if set, then if an error is encountered where the only action possible for an operator is to terminate the **workflow instance**, then the **workflow instance** will be terminated automatically. This avoids the need for the operator to terminate the **workflow instance** manually.

- Concurrent instances: if set, then more than one instance of the workflow context can be running at the same time. This is necessary for batch ingests, where several SIPs are ingested simultaneously, each by its own workflow instance.
- The **Workflow Trigger** option defines how each workflow instance based on this context is started. This can be set to one of 3 values: manual only, scheduled, or start automatically (only available for ingest workflows).
  - When set to **Manual only**, each **workflow instance** must be started by a user via the **Start** tab on the **Ingest** main screen, or from the Actions submenu in Explorer.
  - When set to **Scheduled**, the Preservica system will start a single workflow instance at a regular time as defined in the schedule. When selecting this option the user will need to provide additional information to define when workflow instances should be started. If more than one SIP is found in the source directory, the oldest SIP will be processed.

    Scheduled workflows can be configured to execute:

    - **Daily** - a time (hh:mm) must be specified.
    - **Weekly** - a time (hh:mm) and day of the week (Monday - Sunday) must be specified.
    - **Monthly** - a time (hh:mm) and day of the month (1 - 28) must be specified.
    - **Yearly** - a time (hh:mm), day of the month (1 - 28) and month (January - December) must be specified.
    - **Advanced** - a 5-element CRON schedule must be specified (see http://en.wikipedia.org/wiki/Cron for more details on the CRON format). Note that predefined cron values such as “@daily” are not supported.
  - When set to **Start Ingest Automatically**, the Preservica system will monitor the specified ingest location and start a workflow instance each time a valid ingestable package is found. See the ingest section for details.

The **Workflow Trigger** defining how each workflow instance based on this context is started should not be confused with the fundamental nature of the workflow as determined by the set of workflow steps in the workflow definition. An automatically triggered workflow (whether scheduled or using the file system watcher) could theoretically contain a manual step (a human task); while these workflows will start automatically, they will not complete automatically but will need human intervention. However, typically an automatically triggered workflow will not contain manual steps and so can complete without human intervention.

3.2.2. Managing Workflow Contexts

The **Manage** tab on the workflow management screen displays a list of the *workflow contexts* that have been set up. Specific contexts can be activated or deactivated by ticking the appropriate checkbox in the **Active** column. Active (ticked) workflows are available on the Start tab and, if applicable, on the Actions submenu in Explorer or in search results, and will be started by the system where appropriate. The list also provides access to the following:

- The workflow context definition dialog box.

  This is accessed by clicking on the workflow context name in the *Context Name* column. The dialog box is read only for active workflow contexts, but can be edited if the workflow context is not currently active.

There are some parameters on this screen which are shared between workflows: a name, description, email settings, and the scheduling options at the bottom (see above). See the Standard Workflows document for details of context parameters specific to each workflow definition, which will appear in the middle of this dialog.

- The **Workflow Error Configuration** screen (see Figure 3.5).

  It is displayed by clicking on *Configure Errors* (in the **Errors** column). This screen shows a list of all the configurable errors for the workflow (based on the individual workflow steps within the workflow definition). Some workflow definitions won't specify any configurable errors, and this page will be empty. For each error you can select one of the following actions to be taken:

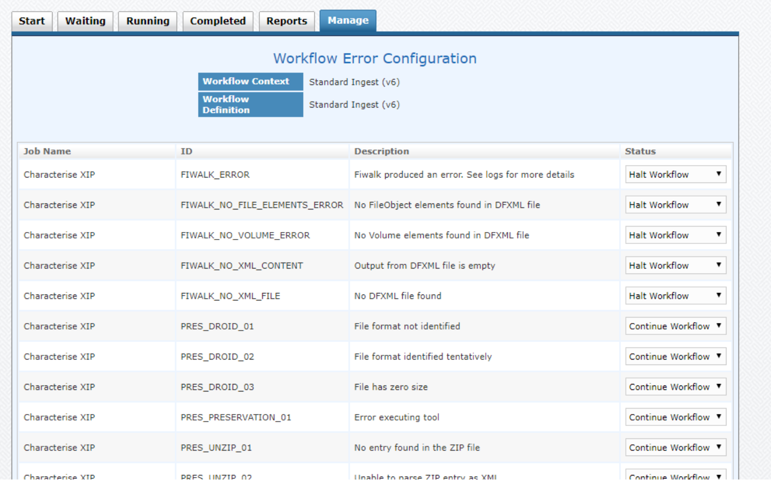
  - **Continue Workflow**: the system will log the error condition but the workflow instance will continue automatically. Generally this setting is used for minor errors or when it is regarded as more important to ingest and secure the content than to resolve errors.

  - **Halt Workflow**: the workflow instance will stop and wait for a user to decide whether to retry or abort the workflow. In order to continue the workflow must support retrying of the step.

  - **Abort Workflow**: the workflow will stop and the only possible user action will be to terminate the workflow instance. Generally this is used for severe errors that would invalidate the system state if the process was allowed to continue, for example being unable to update storage or the database.

Where multiple error conditions are found during the execution of a single workflow step for a specific workflow instance, the action taken by the system will reflect the most severe error condition found.

Figure 3.5. Example workflow error configuration screen



Workflow Error Configuration

Workflow Context	Standard Ingest (v6)
Workflow Definition	Standard Ingest (v6)

Job Name	ID	Description	Status
Characterise XIP	FIWALK_ERROR	Fiwalk produced an error. See logs for more details	Halt Workflow ▾
Characterise XIP	FIWALK_NO_FILE_ELEMENTS_ERROR	No FileObject elements found in DFXML file	Halt Workflow ▾
Characterise XIP	FIWALK_NO_VOLUME_ERROR	No Volume elements found in DFXML file	Halt Workflow ▾
Characterise XIP	FIWALK_NO_XML_CONTENT	Output from DFXML file is empty	Halt Workflow ▾
Characterise XIP	FIWALK_NO_XML_FILE	No DFXML file found	Halt Workflow ▾
Characterise XIP	PRES_DROID_01	File format not identified	Continue Workflow ▾
Characterise XIP	PRES_DROID_02	File format identified tentatively	Continue Workflow ▾
Characterise XIP	PRES_DROID_03	File has zero size	Continue Workflow ▾
Characterise XIP	PRES_PRESERVATION_01	Error executing tool	Continue Workflow ▾
Characterise XIP	PRES_UNZIP_01	No entry found in the ZIP file	Continue Workflow ▾
Characterise XIP	PRES_UNZIP_02	Unable to parse ZIP entry as XML	Continue Workflow ▾

For users with the SDB_MANAGER_USER or SDB_ADMIN_USER roles, an additional **Delete** button is displayed for each context on the Manage tab. Strictly speaking workflow contexts cannot be deleted, in order to maintain the workflow audit trail. However, if a workflow context is marked for deletion, it no longer appears in the list of available workflow contexts on this tab. (Users with the SDB_MANAGER_USER or SDB_ADMIN_USER roles have the option of displaying "deleted" **workflow contexts** by ticking the "Show Deleted Workflows" checkbox under the list.)

Only inactive **workflow contexts** can be deleted. To delete a **workflow context**, click on its "Delete" button, and then click OK on the confirmation dialog. Note that to avoid potential name clash issues, when a **workflow context** is deleted, its name is changed by appending the text `:: Deleted: yyyy-MM-dd hh:mm:ss` (using the actual date / time of deletion).

If showing deleted **workflow contexts**, a user with the SDB_MANAGER_USER or SDB_ADMIN_USER roles will see that the button for any deleted workflows in the **Delete** column is now labelled "Restore". This allows any deleted workflow contexts to be restored back to the system, and so made available once more to general users.

To restore a workflow context, click on its "Restore" button, and then click OK on the confirmation dialog. Note that to avoid potential name clash issues, when a **workflow context** is restored, its name is changed by appending the text `:: Restored: yyyy-MM-dd hh:mm:ss` (using the actual date / time of restoration).

3.2.3. System Workflows

Some workflow definitions are marked as being system workflows. Contexts of these workflow definitions won’t appear for, and can’t be created by, managers.

For administrators, they will also be hidden by default, but there is a check box on the context management pages which will make them visible and editable as normal.

A workflow definition is a system workflow if it has the *constSystemWorkflow* parameter set to true.

Chapter 4. Administration

4.1. Schema Management

This page allows you to manage XML schemas, transforms and documents to configure certain aspects of system behaviour.

4.1.1. XML Schemas

Preservica allows XML schemas (XSD documents) to be registered with the system. Registered schemas can then be used to validate both submissions and uploaded documents for compliance with the schema.

Preservica comes with a number of standard schemas. These include standard archival schemas and schemas used by Preservica for validation purposes:

- The XIP (V6) Schema. XIP is the Preservica import and export metadata schema.
- The XIP (V4) Schema. This is used for ingesting XIP v4 packages (e.g. from the SIP Creator) and some other package formats as an intermediate format.
- The XHTML 1.0 Schema. XHTML is a stricter version of HTML. This schema is used to display metadata for viewing and editing within Explorer.
- The Custom Indexer schema allows the upload of custom search indexers for generic metadata.

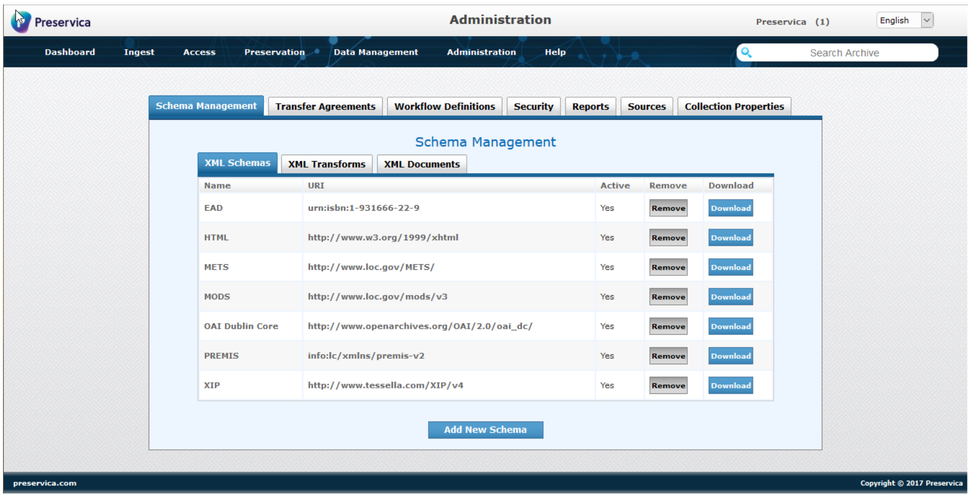
The pre-loaded schemas should not be removed from the system.

The **XML Schemas** tab (see Figure 4.1) displays a list of registered schemas. For each schema the name and URI is shown to aid identification.

- Clicking on the appropriate entry in the **Name** field will display a pop-up screen allowing the user to change the name and description of the schema and a flag to indicate if it is active.
- Clicking on the appropriate entry in the **URI** field will open the URI in a new browser window
- Clicking on the appropriate **Remove** button will, subject to user confirmation, remove the schema from the Preservica system.
- Clicking on the appropriate **Download** button will allow the user to download a copy of the registered schema as a XSD document.

note The list of transforms is paged; you can move between the pages by selecting the appropriate page number below the table on the left-hand side.

Figure 4.1. XML Schema List

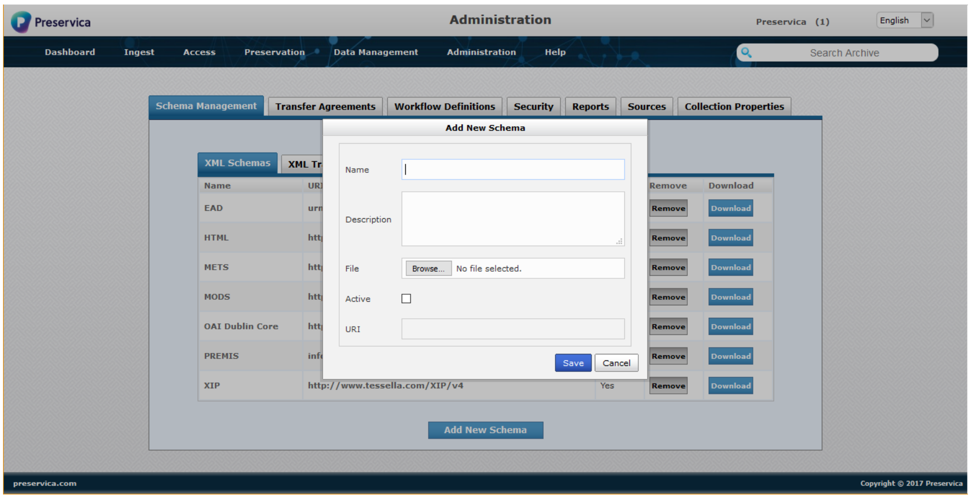


Name	URI	Active	Remove	Download
EAD	urn:isbn:1-931666-22-9	Yes	Remove	Download
HTML	http://www.w3.org/1999/xhtml	Yes	Remove	Download
METS	http://www.loc.gov/METS/	Yes	Remove	Download
MODS	http://www.loc.gov/mods/v3	Yes	Remove	Download
OAI Dublin Core	http://www.openarchives.org/OAI/2.0/oai_dc/	Yes	Remove	Download
PREMIS	info:lc/xmlns/premis-v2	Yes	Remove	Download
XIP	http://www.tessella.com/XIP/v4	Yes	Remove	Download

The user can also add a new schema to the system by pressing the **Add New XSD Schema** button. This will open an **Add New XSD Schema** dialog box (see Figure 4.2) where the following information should be entered:

- The **Name** of the schema for display and reference purposes. This entry is mandatory and must be unique.
- A free text **Description** of the schema or its purpose. This entry is mandatory.
- The location of the schema definition (XSD) file; which must be on a locally accessible drive. A browse window will open to help locate this file.
- A flag to indicate if the schema is active.

Figure 4.2. Add XML Schema



Add New Schema

Name

Description

File Browse... No file selected.

Active □

URI

The schema can be added by clicking on the **save** button.

The URI (Uniform Resource Identifier) is the address associated with the XSD schema as retrieved from the XML **targetNamespace** element within the XSD file. Preservica will not allow files without this XML element to be uploaded.

4.1.2. XML Transforms

Preservica allows XML Transforms (XSLT documents) to be registered with the system. Registered transforms can then be used to convert the format of XML metadata held within Preservica to another XML schema.

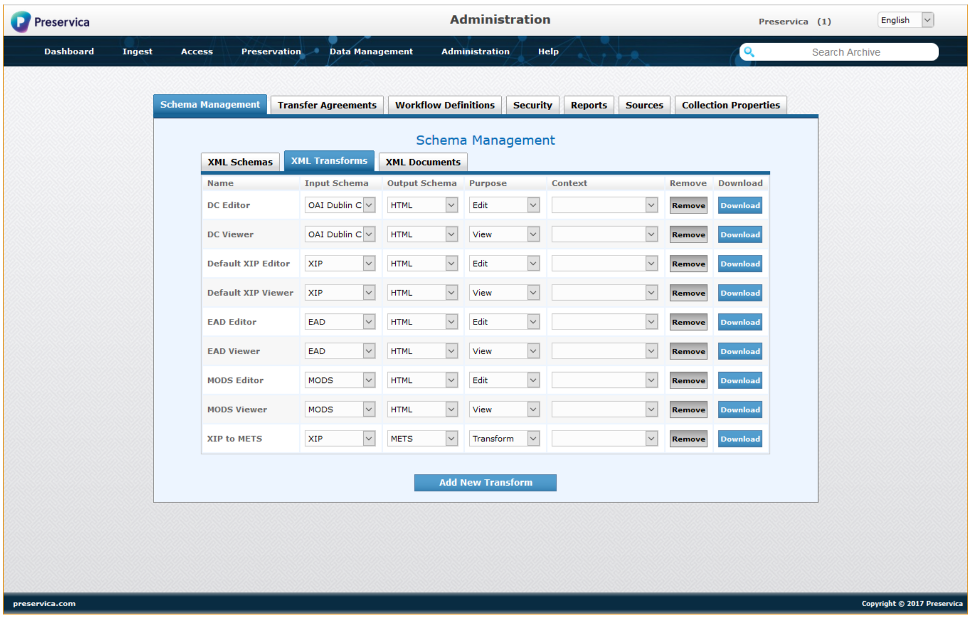
The **XML Transforms** tab (see Figure 4.3) displays a list of registered transforms. For each transform the name and the input and output schemas is shown together with the purpose and context of the transform.

- Clicking on the appropriate entry in the **Name** field will display a pop-up screen allowing the user to change the name of the Transform.
- The appropriate drop down list in the **Input Schema** column will allow the user to select the format of the source XML to which this transform applies from a list of registered schemas.
- The appropriate drop down list in the **Output Schema** column will allow the user to select the format of the target XML to which this transform applies from a list of registered schemas.
- The appropriate drop down list in the **Purpose** column will allow the user to select the purpose (Edit, View or Transform) of the transform.
  - Edit transforms are used in Explorer (on the Properties screen) to allow users with the appropriate privileges to edit metadata that conforms to the schema listed in the XSD Input column. To edit metadata in Explorer, the schema in the Output Schema column **must** be set to XHTML 1.0 or left blank.
  - View transforms allow users to view metadata (on the Properties screen of Explorer) that conforms to the schema listed in the Input Schema column. To view metadata in Explorer, the schema in the Output Schema column **must** be set to XHTML 1.0 or left blank.
  - Transforms with a purpose of *Transform* are used to transform XML metadata between the given pair of schemas. Typically this is used on ingest or access to transform to or from the XIP metadata schema, but it can be used to transform between any pair of schemas. The other common use of the Transform purpose is to transform fragments to the CMIS Metadata schema to make metadata available through the Content API and to Universal Access.
- The appropriate drop down list in the **Context** column will allow the user to optionally differentiate transforms with the same Input, Output and Purpose based upon the context in which they are used (Ingest, Access, Data Management, Preservation).
- Clicking on the appropriate **Remove** button will, subject to user confirmation, remove the transform from the Preservica system.
- Clicking on the appropriate **Download** button will allow the user to download a copy of the registered transform as a XSLT document.

note

The list of transforms is paged; you can move between the pages by selecting the appropriate page number below the table.

Figure 4.3. XML Transform List



Name	Input Schema	Output Schema	Purpose	Context	Remove	Download
DC Editor	OAI Dublin C	HTML	Edit		Remove	Download
DC Viewer	OAI Dublin C	HTML	View		Remove	Download
Default XIP Editor	XIP	HTML	Edit		Remove	Download
Default XIP Viewer	XIP	HTML	View		Remove	Download
EAD Editor	EAD	HTML	Edit		Remove	Download
EAD Viewer	EAD	HTML	View		Remove	Download
MODS Editor	MODS	HTML	Edit		Remove	Download
MODS Viewer	MODS	HTML	View		Remove	Download
XIP to METS	XIP	METS	Transform		Remove	Download

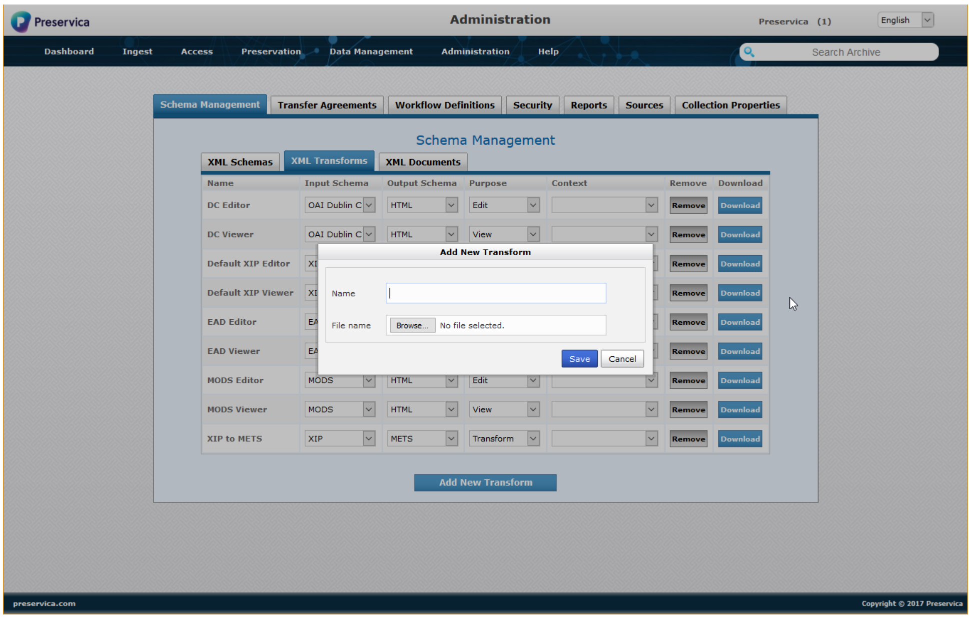
The user can also add a new transform to the system by pressing the **Add New XSLT Transform** button. This will open an **Add New XSLT Transform** dialog box (see Figure 4.4) where the following information should be entered:

- The **Name** of the Transform for display and reference purposes. This entry is mandatory and must be unique.
- The location of the transform definition (XSLT) file which should be on the local machine. A browse window will open to help locate this file.

Once the transform has been added (by pressing the **Save** button) it will be listed on the **XML Transforms** tab. At this point the source and target XML schemas to which the transform applies and the transform's purpose (and optionally its context) can be defined. For View and Edit transforms the target XML schema can be left blank or the XHTML schema selected.

Generic view and edit transforms which display any schema in a simple name/value pairs are available. If these transforms are to be used the source XML schema must be left blank.

Figure 4.4. Add New XML Transform



Name	Input Schema	Output Schema	Purpose	Context	Remove	Download
DC Editor	OAI Dublin C	HTML	Edit		Remove	Download
DC Viewer	OAI Dublin C	HTML	View		Remove	Download
Default XIP Editor	XI				Remove	Download
Default XIP Viewer	XI				Remove	Download
EAD Editor	EA				Remove	Download
EAD Viewer	EA				Remove	Download
MODS Editor	MODS	HTML	Edit		Remove	Download
MODS Viewer	MODS	HTML	View		Remove	Download
XIP to METS	XIP	METS	Transform		Remove	Download

Add New Transform

Name

File name Browse... No file selected.

4.1.3. XML Documents

Preservica allows XML documents to be registered with the system; these documents are valid fragments of XML. Registered documents can then be used within Preservica for one of the following purposes depending upon the value set for their “Context”:

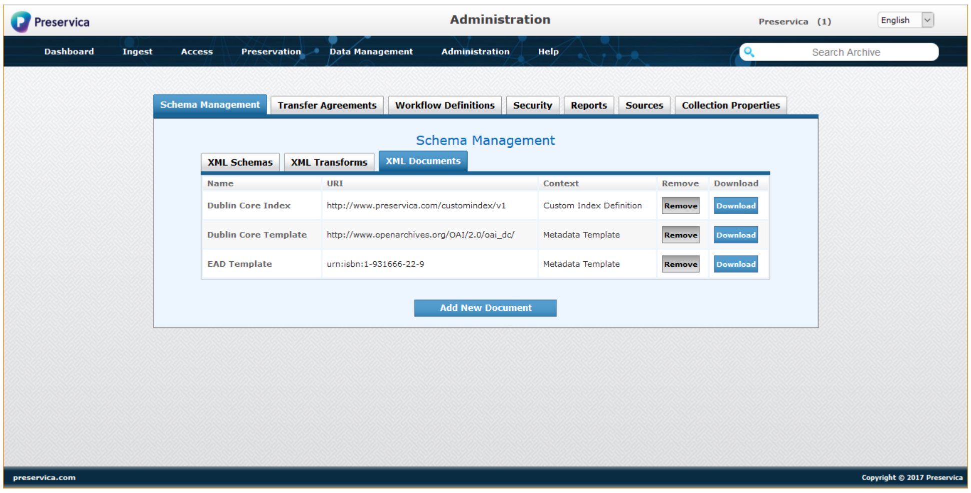
- **Metadata Template** XML documents can be used within Explorer to provide a standard template to add blocks of additional metadata to an entity (folder or asset)
- **Configuration File** XML documents can be used to specify part of the configuration for a workflow. Currently two types of configuration file can be registered with the system.
    - Documents that conform to the CatalogueMapping.xsd schema are used to customise the synchronisation of information between an external catalogue or archival information system (AIS), such as Axiell’s CALM, and the metadata store.
    - Documents that conform to the heritrix_settings.xsd schema are used to customise the behaviour of a website harvest performed by the Heritrix web crawler used to ingest websites.
- **Custom Index Definition** documents define a custom search indexer for generic metadata. These documents should fit the Custom Indexer schema. See Chapter 10 for details of setting up custom search fields for metadata.
- **Metadata Bulk Edit List** documents define the fields of metadata available for bulk edit operations. Examples of how to write these documents can be found in the released xml template files. These documents should fit the bulk edit list schema, and the following rules will apply:
    - Each document points to a single Metadata Schema, specified by the schemaUri tag.
    - Any number of fields can be made editable by including them with the editTerm tag.
    - Fields must be indexed in order to be editable.
    - Each field can made be editable for all or a subset of the types of edits that are available [Update,Replace,Clear,Remove] by using the editType tag. (Add Fragment does not require white listing.)

- Each field can made be editable (white-listed) or not editable (black-listed) for the editType's specified by using the editable tag.
- If multiple editTerms exist for the same field, any black-list declarations will take priority over the corresponding white-list declaration.

The **XML Document** tab (see Figure 4.5) displays a list of registered documents. For each document its name, the underlying schema URI, and its context is shown.

- Clicking on the appropriate entry in the **Name** field will display a pop-up dialog box allowing the user to change the name, URI and context of the document.
- Pressing the appropriate **Remove** button will, subject to user confirmation, remove the document from the Preservica system.
- Pressing the appropriate **Download** button will allow the user to download a copy of the registered document as a XML document.

Figure 4.5. XML Document List



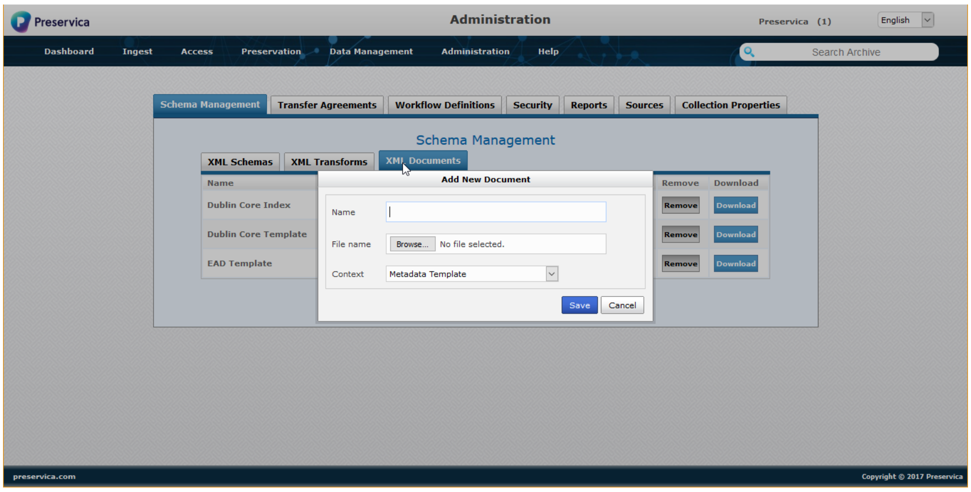
Name	URI	Context	Remove	Download
Dublin Core Index	http://www.preservica.com/customindex/v1	Custom Index Definition	Remove	Download
Dublin Core Template	http://www.openarchives.org/OAI/2.0/oai_dc/	Metadata Template	Remove	Download
EAD Template	urn:isbn:1-931666-22-9	Metadata Template	Remove	Download

The user can also add a new document to the system by pressing the Create **New XML Document** button. This will open a **Create New XML Document** dialog box (see Figure 4.6) where the following information should be entered:

- The **Name** of the document for display and reference purposes. This entry is mandatory and must be unique.
- The location of the document XML file which should be on the local machine. A browse window will open to help locate this file.
- The **Context** in which the document is to be used. This defaults to **Metadata Template**, which is used for adding metadata fragments to entities in Explorer.

The XML file specified must be valid (well formed) and must contain a schema URI (Uniform Resource Identifier) this is the address associated with the underlying XSD schema as retrieved from the XML **targetNamespace** element within the XSD file. Preservica will not allow files without this XML element to be uploaded. Also, Preservica will check that the underlying schema is registered and if it is not, will prevent the XML document from being registered.

Figure 4.6. Add XML Document



Name	Remove	Download
Dublin Core Index	Remove	Download
Dublin Core Template	Remove	Download
EAD Template	Remove	Download

Name:

File name: Browse... No file selected.

Context: Metadata Template

4.2. Security

User access rights and permissions within Preservica are mapped to user roles, which are assigned and managed either in an external system, such as Microsoft Active Directory and OpenLDAP, or using the built in Preservica User Management module (see Section 4.9 for more details).

When a user logs in to the system, Preservica will authenticate the user against the user credentials store, and obtain the user's details (name, e-mail address, Organisational Unit) as well as the roles that have been assigned to that user.

System security can be applied by two independent mechanisms. Firstly, access to system functionality is controlled using a set of pre-configured Preservica roles (see Section 4.2.1), and secondly, access to content is controlled using permissions associated with either the pre-configured Preservica roles or by the addition of custom roles (see Section 4.2.2).

Any additional roles that are added to the system only govern access to content and metadata and do not affect a user's ability to access system functions.

Preservica provides a default configuration of content security using the pre-configured functional roles and two access control tags ("open" and "closed").

4.2.1. Access control by high level function

Preservica is divided into "functional areas" that broadly correspond to the functional entities of the OAIS reference model (and the links in the page header). A set of fixed functional roles exist that provide access to those areas:

- **ROLE_SDB_ACCESS_USER -** This role allows a user to browse and retrieve files and DIPs.
- **ROLE_SDB_SUBMITTER_USER -** The role only allows access to page which lets the user download the upload wizard installer. A user with this role will have no access to any other features of Preservica.
- **ROLE_SDB_ADMIN_USER -** This role grants a user access to the administration functionalities of the Preservica application, and full access to the Job Queue and Explorer applications. Caution should be exercised in granting this role. An unwitting or malicious user with admin privileges could cause serious damage to the archive infrastructure. This role is not available to you on our cloud hosted systems.
- **ROLE_SDB_ANONYMOUS_USER -** This role is used by the Universal Access application to control the content and metadata shown to unauthenticated users. It doesn't have any functional permissions in the application.

- **ROLE_SDB_DATA_MANAGEMENT_USER -** This role grants a user access to the data management workflow system in Preservica, which allows a user to re-index, re-characterise, appraise, delete, and generally manage the archive's content.

- **ROLE_SDB_INGEST_USER -** This role grants a user access to the ingest workflows system in Preservica. This allows them to submit data to the archive.

- **ROLE_SDB_MANAGER_USER -** This grants a user access to the archive administration functionalities of the Preservica application, and full access to the Job Queue and Explorer applications. Caution should be exercised in granting this role. An unwitting or malicious user with manager privileges could cause serious damage to the archive's data.

- **ROLE_SDB_REGISTRY_ADMIN_USER -** A user with this role can edit existing entries in the Registry and create new ones. This role is not available in Preservica CE.

- **ROLE_SDB_TRANSFORM_USER -** This grants a user access to the transformation workflows system in Preservica. This allows them to create and schedule format migrations, creating new accessions both for archival and for testing purposes.

These fixed roles come pre-configured within Preservica. In the case of an on-site Preservica installation they must be added to any external authentication system. Further, these are the only roles in Preservica that can be used to control access to these functional areas, any custom roles can only define content security. This means that if a user is required to be able to run ingest workflows they must be granted one of the `ROLE_SDB_ADMIN_USER`, `ROLE_SDB_MANAGER_USER` or `ROLE_SDB_INGEST_USER` roles. Similarly, to be able to run ingest and preservation workflows a user must have either one of `ROLE_SDB_ADMIN_USER` or `ROLE_SDB_MANAGER_USER`, or both `ROLE_SDB_INGEST_USER` *and* `ROLE_SDB_TRANSFORM_USER`.

Within the user management system the required user accounts should be identified (or created) and the appropriate functional role (Security Groups in Active Directory) or roles assigned to allow the users to access Preservica functions.

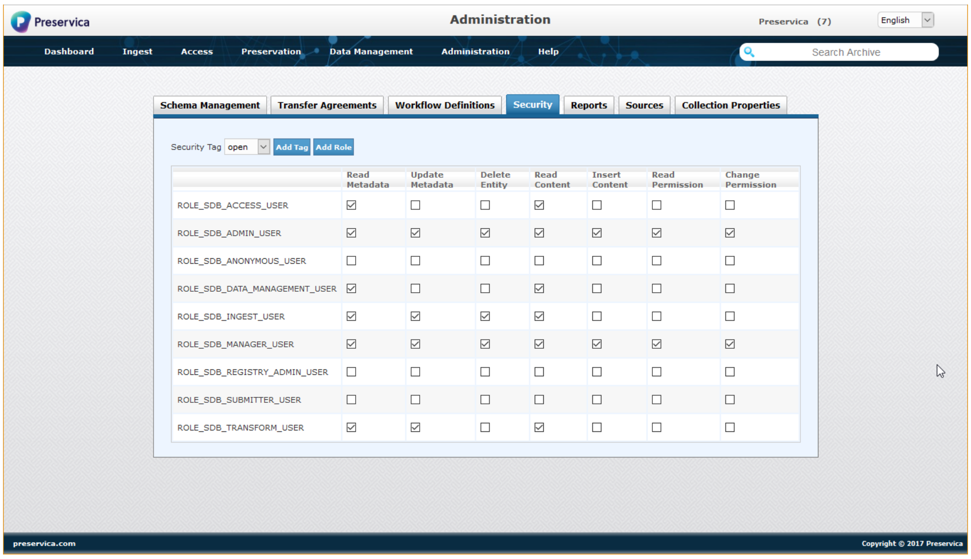
4.2.2. Content Security

Preservica supports the concept of security classification of content, e.g. Open, Restricted, Secret, Top Secret etc., via the XIP metadata field "SecurityDescriptor". If a user does not have access to the appropriate security classification they cannot see the content and/or metadata and it will not be returned by search activities etc.

Preservica has a flexible approach to implementing a security classification scheme. The system is initially configured with two security classification levels, 'open' (visible to all users) and 'closed' (not visible to any users). Preservica treats all security classifications as independent and does not hold a hierarchy of classification levels.

The default configuration for the 'open' setting is shown in Figure 4.7. For more details on the permissions see Section 4.2.3 below.

Figure 4.7. Default security configuration



Security Tag open

	Read Metadata	Update Metadata	Delete Entity	Read Content	Insert Content	Read Permission	Change Permission
ROLE_SDB_ACCESS_USER	☑	□	□	☑	□	□	□
ROLE_SDB_ADMIN_USER	☑	☑	☑	☑	☑	☑	☑
ROLE_SDB_ANONYMOUS_USER	□	□	□	□	□	□	□
ROLE_SDB_DATA_MANAGEMENT_USER	☑	□	□	☑	□	□	□
ROLE_SDB_INGEST_USER	☑	☑	☑	☑	□	□	□
ROLE_SDB_MANAGER_USER	☑	☑	☑	☑	☑	☑	☑
ROLE_SDB_REGISTRY_ADMIN_USER	□	□	□	□	□	□	□
ROLE_SDB_SUBMITTER_USER	□	□	□	□	□	□	□
ROLE_SDB_TRANSFORM_USER	☑	☑	□	☑	□	□	□

Additional security tags can be created as required: press the **Add Tag** button and enter the name of the new security tag in the dialog box that appears. The new security tag(s) can be used to apply different levels of access to archival content. Created tags will be added to the drop-down list of security tags and by selecting the security tag, the appropriate permissions for the different roles can be set in the table.

Additional *content* security roles can be configured as required: press the **Add Role** button and enter the name of the new security role in the dialog box that appears. Any *new* security roles will have no functional permissions. The new role(s) will be added to the list of roles and the appropriate permissions for the different security tags can be set. Roles added to Preservica must be prefixed with the characters “ROLE_” (omitting the double quotes). Before these roles can be assigned to an individual user, they must be added to the user management system where the “ROLE_” prefix should *not* be included.

4.2.3. Security Model

Preservica defines seven distinct permissions, which can be set on a “per tag, per role” basis. For a given tag, these permissions are defined below:

- **Read Metadata** - Users with roles that are granted this permission are able to see the metadata associated with archived entities. In particular, this means that they can browse Explorer for these entities and see them in API results. Users without Read Metadata permission won’t see that an entity exists at all.
- **Update Metadata** - Users with roles that are granted this permission are able to change the metadata associated with archived entities. Creating new entities, links etc are considered to be metadata changes and so this permission also allows users to perform these operations. All changes are audited.
- **Delete Entity** - Users with roles that are granted this permission are able to run Delete workflows on archived entities, removing them completely from the archive.
- **Read Content** - Users with roles that are granted this permission are able to see the digital content, i.e. the digital files, of archived entities. In particular, they are able to download individual files (through Explorer or APIs), or include those files in DIPs created through export workflows.
- **Insert Content** - Users with roles that are granted this permission are able to store content during ingest. This permission is also required to save new content generated by the system during migration.
- **Read Permission** - Users with roles that are granted this permission are able to read the security tag for archived entities. This can be displayed in Preservica Explorer through the entity context menu.

- **Change Permission** - Users with roles that are granted this permission are able to change the security tag for archived entities. This can be set in Preservica Explorer through the entity context menu (see the system user guide [SUG] for more details).

The check boxes on the ‘sdb/security.html’ page allow for live updating of these permissions for each security tag (i.e. the permissions take effect immediately).

Whether an individual user can carry out a specific action depends on the combination of their functional role(s) and content security permissions. That is, the user needs to have an appropriate functional role (see ???) in order to access the required functional area within Preservica, but in addition the user needs to have the appropriate content permission(s) for the archival entities affected by the action.

4.2.3.1. Content Security and Ingest

When content is ingested the security classification of the content (as defined by the “SecurityTag” field in the XIP metadata) is checked. The ingesting user[1] must have the Insert Content permission on the security descriptor assigned to every Information Object and Content Object in the ingest package, and must have Update Metadata permission on the security descriptor assigned to every Structural Object (folder) in the package, and also on the folder into which the package is being ingested.

Even if the user performing the ingest has access to the Preservica Explorer application, they will not be able to see any information relating to the ingested content unless they have the correct security permissions (i.e. Read Metadata permission).

4.2.3.2. Content Security required for Preservation

To run a preservation workflow against existing archival entities, a user needs both a suitable role to access Explorer (either ROLE_SDB_ADMIN_USER, ROLE_SDB_MANAGER_USER, or ROLE_SDB_ACCESS_USER) and specific content permissions on *all* the content being transformed: **read metadata** and **read content**. Read metadata permission is required so the user can select the content to be preserved, while read content permission is required as the content can be viewed by the user if a test transformation is undertaken.

To save the results of the migration, the user will also need the same content permissions as for ingest, i.e. **insert content** and **update metadata**.

4.2.3.3. Example Scenarios

This section provides some scenarios for configuring the security sub-system and is provided for guidance only.

Archive with simple content security requirements

Preservica is initially configured with two security classifications: “open” (visible to all users) and “closed” (not visible to any users). By default the six main functional roles have content security set to enable access to open records as shown in Table 4.1.

Neither the registry administrator role (`ROLE_SDB_REGISTRY_ADMIN_USER`) nor the anonymous access role have any functional area access rights within the Preservica application.

Table 4.1. Default permissions for the “open” security classification

	Read Metadata	Update Metadat	Delete Entity	Read Content	Insert Content	Read Permissi	Change Permissi
Access	y			y			

[1] For a workflow started manually, this is the user that started it. For a scheduled workflow, this is the user who created the workflow context.

	Read Metadata	Update Metadat	Delete Entity	Read Content	Insert Content	Read Permissi	Change Permissi
Admin	y	y	y	y	y	y	y
Anonymous							
Data management	y			y			
Ingest	y	y	y	y			
Manager	y	y	y	y	y	y	y
Transform	y	y		y			
Registry admin							
Submitter							

The default content security has no permissions set for any user roles on the ‘closed’ tag. If it proves necessary to access closed records the necessary operations for the closed tag will need to be assigned to a new or existing role.

Where this simple security model is sufficient, the ability to combine the functional and content security roles reduces the configuration required in the external authentication system.

Archive requiring a content security hierarchy

Where a more complex security classification system is required this can be provided in a number of different ways, but will require configuration of both Preservica and the external authentication system.

For example, if we assume the following hierarchical content classification is required:

- Closed
- Top Secret
- Secret
- Restricted
- Open

where open records can be accessed by all users of the archive and closed records cannot be seen by any users of the archive.

As Preservica is not aware of hierarchical security permissions the hierarchy will need to be built up either within Preservica or within the Authentication Service.

In the following example a number of groups are set up within the Authentication Service to map onto the content security classification. These roles are then reflected in Preservica; note that any roles set up within Preservica must have a prefix of “ROLE_” added to the name set within the Authentication Service. Within Preservica each role must be granted the appropriate permissions for the corresponding content security “tag”. In this approach, a user given the role SDB_SECRET will have access to records with a security tag of secret but not higher or lower classifications (open, restricted, etc.). To provide access to the lower levels of content security it is necessary to assign the corresponding roles to the user within the Authentication Service, effectively building the hierarchy of the security classifications within the Authentication Service. While this approach has the advantage of simplifying the mapping of roles to content security tags within Preservica, it will require a number of roles to be assigned to each user within the Authentication Service unless the Authentication Service supports establishing groups of groups.

Content Security Required	Authentication System Role	Preservica Role Required	Linked security tags (within Preservica)
Open	SDB_ACCESS_USER SDB_ACCESS_OPEN	ROLE_SDB_ACCESS_USER ROLE_SDB_ACCESS_OPEN	None open
Restricted	SDB_ACCESS_USER SDB_ACCESS_OPEN SDB_ACCESS_RESTRICTED	ROLE_SDB_ACCESS_USER ROLE_SDB_ACCESS_OPEN ROLE_SDB_ACCESS_RESTRICTED	None open restricted
Secret	SDB_ACCESS_USER SDB_ACCESS_OPEN SDB_ACCESS_RESTRICTED SDB_ACCESS_SECRET	ROLE_SDB_ACCESS_USER ROLE_SDB_ACCESS_OPEN ROLE_SDB_ACCESS_RESTRICTED ROLE_SDB_ACCESS_SECRET	None open restricted secret
Top Secret	SDB_ACCESS_USER SDB_ACCESS_OPEN SDB_ACCESS_RESTRICTED SDB_ACCESS_SECRET SDB_ACCESS_TOP_SECRET	ROLE_SDB_ACCESS_USER ROLE_SDB_ACCESS_OPEN ROLE_SDB_ACCESS_RESTRICTED ROLE_SDB_ACCESS_SECRET ROLE_SDB_ACCESS_TOP_SECRET	None open restricted secret top secret
Closed	None	None	None

To minimise the number of roles that need to be assigned to each user within the Authentication Service it is possible to build the hierarchy of the security classification within Preservica as shown in the following example. This approach will not reduce the number of roles (groups) that need to be set up in the Authentication Service but will reduce the number of roles assigned to each user.

Content Security Required	Authentication System Role	Preservica Role Required	Linked security tags (within Preservica)
Open	SDB_ACCESS_USER SDB_ACCESS_OPEN	ROLE_SDB_ACCESS_USER ROLE_SDB_ACCESS_OPEN	None open
Restricted	SDB_ACCESS_USER SDB_ACCESS_RESTRICTED	ROLE_SDB_ACCESS_USER ROLE_SDB_ACCESS_RESTRICTED	None open, restricted
Secret	SDB_ACCESS_USER SDB_ACCESS_SECRET	ROLE_SDB_ACCESS_USER ROLE_SDB_ACCESS_SECRET	None open, restricted, secret
Top Secret	SDB_ACCESS_USER SDB_ACCESS_TOP_SECRET	ROLE_SDB_ACCESS_USER ROLE_SDB_ACCESS_TOP_SECRET	None open, restricted, secret, top secret
Closed	None	None	None

If required this pattern can be simplified by associating the lowest security classification content security with the built-in functional roles. However, this does blur the distinction between content and functional roles.

Content Security Required	Authentication System Role	Preservica Role Required	Linked security tags (within Preservica)
Open	SDB_ACCESS_USER	ROLE_SDB_ACCESS_USER	open
Restricted	SDB_ACCESS_USER SDB_ACCESS_RESTRICTED	ROLE_SDB_ACCESS_USER ROLE_SDB_ACCESS_RESTRICTED	open restricted
Secret	SDB_ACCESS_USER SDB_ACCESS_SECRET	ROLE_SDB_ACCESS_USER ROLE_SDB_ACCESS_SECRET	open restricted, secret
Top Secret	SDB_ACCESS_USER SDB_ACCESS_TOP_SECRET	ROLE_SDB_ACCESS_USER ROLE_SDB_ACCESS_TOP_SECRET	open restricted, secret, top secret
Closed	None	None	None

4.3. Reports

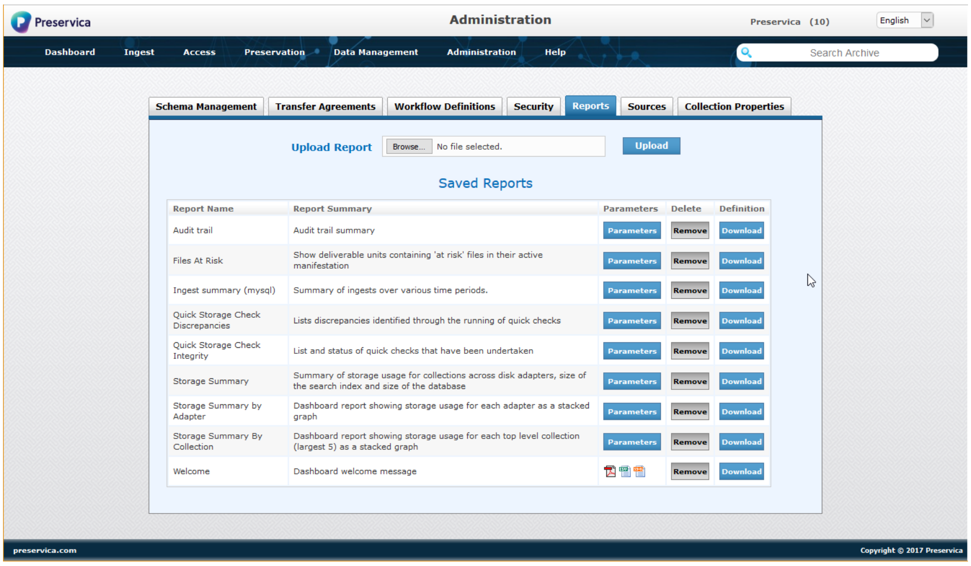
Preservica utilises the Jasper Reports [http://community.jaspersoft.com/] framework to provide an in-built reporting capability. Preservica is delivered with a set of standard reports; these reports are detailed in Preservica Standard Reports [SRP]. Additional reports can be created using the iReport [http://community.jaspersoft.com/project/ireport-designer] graphical report designer and uploaded. All reports should contain a “Report Type” indicating which area of the system they relate to, but all loaded reports are visible within the administration function.

The **Reports** tab (see Figure 4.8) displays a list of loaded reports. For each report the name and description are shown together with options to run the report (specifying any parameters that are required), download or remove it.

note

Uploading, downloading and removing report definitions is only available to users with the SDB_ADMIN_USER role.

Figure 4.8. Report List



Report Name	Report Summary	Parameters	Delete	Definition
Audit trail	Audit trail summary	Parameters	Remove	Download
Files At Risk	Show deliverable units containing 'at risk' files in their active manifestation	Parameters	Remove	Download
Ingest summary (mysql)	Summary of ingests over various time periods.	Parameters	Remove	Download
Quick Storage Check Discrepancies	Lists discrepancies identified through the running of quick checks	Parameters	Remove	Download
Quick Storage Check Integrity	List and status of quick checks that have been undertaken	Parameters	Remove	Download
Storage Summary	Summary of storage usage for collections across disk adapters, size of the search index and size of the database	Parameters	Remove	Download
Storage Summary by Adapter	Dashboard report showing storage usage for each adapter as a stacked graph	Parameters	Remove	Download
Storage Summary By Collection	Dashboard report showing storage usage for each top level collection (largest 5) as a stacked graph	Parameters	Remove	Download
Welcome	Dashboard welcome message	PDF CSV XML	Remove	Download

- Where the report does not require any input parameters small icons will be displayed allowing the report to be produced in PDF, CSV or XML format.

- Where input parameters are required for a report, a single **Parameters** button is displayed in the Parameters column. Pressing this button will display an additional dialog box (see Figure 4.9) with the appropriate input parameter fields for the selected report. Once the required parameters have been entered, the report can be produced in the required format by clicking on the appropriate icon. The following default values apply:

  - Text fields default to "%", the wildcard which will match all values.

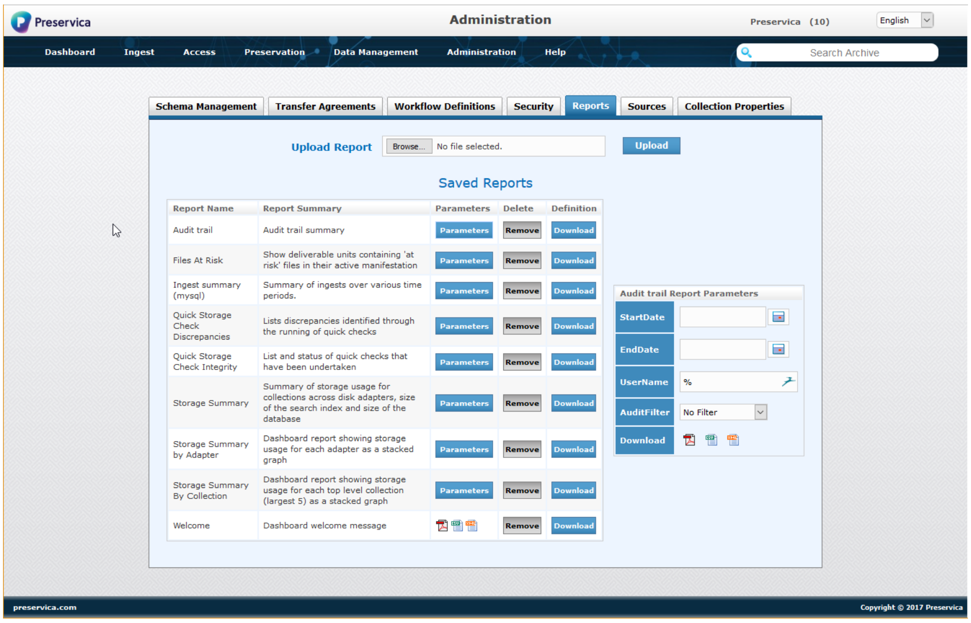
  - The start date for a date range defaults to blank, which means that no restriction will be applied based on the start date, other than it must be before the end date.

  - The end date for a date range defaults to blank, which means that no restriction will be applied based on the end date, other than it must be after the start date.

To add a new report to Preservica, press the **Browse…** button, which will open a browse window to allow you to locate and select the file to upload. The report definition (JRXML) file must be located on a drive accessible to your web browser. Once you have selected your report, press the **Upload** button to upload it to Preservica. Preservica will validate the syntax of the report definition and reject any invalid files.

Reports can be scheduled to be produced automatically on a Daily, Weekly, Monthly or Yearly basis by using the Send Automated Reports data management workflow. Which report is set, when, in what format and to which email address can be specified in the workflow context definition.

Figure 4.9. Report Parameters



Saved Reports

Report Name	Report Summary	Parameters	Delete	Definition
Audit trail	Audit trail summary	Parameters	Remove	Download
Files At Risk	Show deliverable units containing 'at risk' files in their active manifestation	Parameters	Remove	Download
Ingest summary (mysql)	Summary of ingests over various time periods.	Parameters	Remove	Download
Quick Storage Check Discrepancies	Lists discrepancies identified through the running of quick checks	Parameters	Remove	Download
Quick Storage Check Integrity	List and status of quick checks that have been undertaken	Parameters	Remove	Download
Storage Summary	Summary of storage usage for collections across disk adapters, size of the search index and size of the database	Parameters	Remove	Download
Storage Summary by Adapter	Dashboard report showing storage usage for each adapter as a stacked graph	Parameters	Remove	Download
Storage Summary By Collection	Dashboard report showing storage usage for each top level collection (largest 5) as a stacked graph	Parameters	Remove	Download
Welcome	Dashboard welcome message	📄 CSV XML	Remove	Download

Audit trail Report Parameters

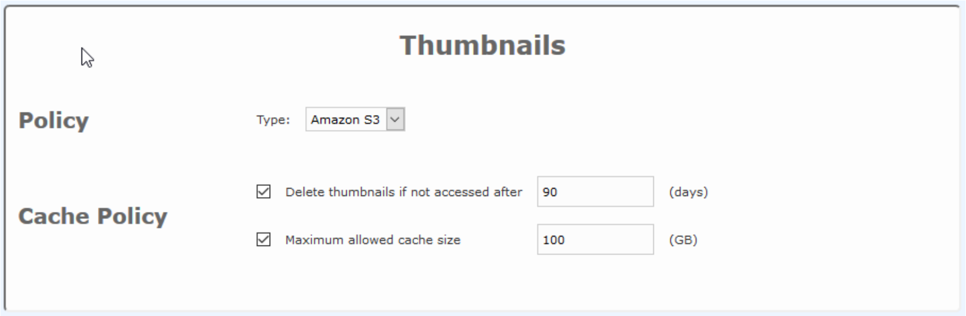
StartDate	
EndDate	
UserName	%
AuditFilter	No Filter
Download	📄 CSV XML

4.4. Collection Properties

Collection Properties provide a means of configuring the system behaviour for the current tenancy. Changes to these properties will not affect the behaviour of other tenancies.

4.4.1. Thumbnails

Figure 4.10. Thumbnail Policies



Thumbnails

Policy

Type: Amazon S3

Cache Policy

☑ Delete thumbnails if not accessed after 90 (days)

☑ Maximum allowed cache size 100 (GB)

This part of the Collection Properties tab stores the configuration settings for thumbnailing. This includes setting thumbnail storage policy and cache policy.

The policy offers a "Type" drop-down menu, which will offer, depending on your configuration

- Off - Do not generate or store new thumbnails. Setting your policy to none does not remove any existing thumbnails.
- Database - Store thumbnails in the Preservica database. This is not recommended for production use as it can cause unnecessary growth of the database.
- Any Sources (see Section 4.7) configured as Thumbnail Storage. On our hosted systems there is likely to be an S3 or Azure option.

Cache Policy allows you to control the amount of storage used by thumbnails, either by setting a maximum cache size (in GB), or by explicitly asking the system to delete thumbnails that have not been accessed for some period of time (specified in days), or both. If a maximum cache size is set, once this limit has been reached, newly created thumbnails continue to be added, but the system will automatically remove older thumbnails to bring the cache size back down. The system removes those thumbnails whose last accessed date is furthest in the past.

Thumbnails for content stored on off-line adapters, such as Amazon Glacier, are not removed from the cache due to the cost of recreating these at a later point.

You can clear your thumbnail cache by setting your policy to Off AND by selecting *Delete thumbnails if not accessed after* and setting the value to 0 days. The cache will be cleared by the system at some point in the next 24 hours.

If your Thumbnail policy is not set to None, thumbnails are automatically generated for individual assets as a background task after ingest. Additionally, attempting to display the thumbnail for a file either through Explorer, or the Universal Access system will automatically generate a thumbnail if one does not currently exist (for example, if thumbnailing failed during the ingest process). Thumbnails can be manually set for entities through the Explorer interface (see [SUG] for details). These thumbnails are also not removed from the cache if the system needs to manage the cache size.

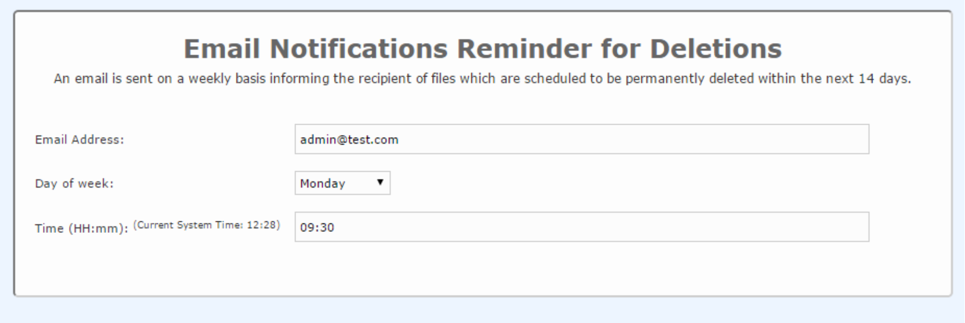
The automatic thumbnailing settings let you control whether thumbnails are automatically selected for folders. If you turn this on, folders which have no thumbnail will have one automatically assigned to them, from one of the assets within that folder. Any asset within the folder which already has a thumbnail is eligible, but if there are multiple possibilities, the system will pick one arbitrarily. If you set this to None, folders which don’t have a thumbnail explicitly assigned to them will have no thumbnail. Automatic thumbnailing only applies to direct descendants, i.e. folders with assets directly contained within them; it will not look within child folders to locate an asset for a thumbnail.

Default thumbnails are used if an item would otherwise have no thumbnail. This means assets that cannot be thumbnailed (for example because they are in an unknown format, or stored in an inaccessible or off-line location), and any folders for which automatic thumbnailing does not apply and for which no specific thumbnail has been selected in Explorer. The thumbnail selected here will be shown in Explorer, Universal Access and through the Content API as if it was the thumbnail for the item in question.

The thumbnail style allows you to control how Preservica generates thumbnails when the input is not square. Note that changing this setting only applies to newly generated thumbnails; it won’t affect any existing ones. The options are:

- Boxed: The image will be resized, preserving aspect ratio, and placed in the centre of a box such that the box is of the requested size. For example, if creating a 400×400 preview from a 2000×1000px image, the thumbnail will be a 400×200 version of the image centred within a 400×400px box.
- Stretched: The image will be resized to fit the preview, ignoring aspect ratio.
- Cropped: The image will be resized, preserving aspect ratio, and a square section of the requested size will be taken from the resized image.

4.4.2. Permanent Deletion Email Reminders



Email Notifications Reminder for Deletions

An email is sent on a weekly basis informing the recipient of files which are scheduled to be permanently deleted within the next 14 days.

Email Address: admin@test.com

Day of week: Monday

Time (HH:mm): (Current System Time: 12:28) 09:30

A notification email will be sent on a weekly basis to inform a nominated user the number of folders and assets which are scheduled to be permanently deleted from the system within the next 14 days.

To configure the day of the week, time of day, and the receiving email address, set the corresponding values in the panel shown above and then click **Save Changes**.

If any files are suspected to have been accidentally deleted, it is possible to recover the content before the content is permanently deleted. Please see the *Recoverable Deletion* section (Section 4.8) for more information.

4.4.3. Content Metadata

This section allows you to choose which metadata fields will be used for the Title and Description when requesting object details via CMIS or the Content API (and therefore also Universal Access, which uses the Content API). Select the field which you want to use for each of those two things.

If you have a custom XIP to CMIS transform uploaded and active, these settings won't be applied.

The responses to those APIs are cached, so when you change your selection here, it may not be applied immediately for all entities.

4.4.4. Tenant System Properties

The Collection Properties page also contain additional Tenant System Properties which can be individually configured per tenant. For System Properties which are configured system-wide, and which apply to all tenants, please see section Section 5.3.

Property Name	Notes
extract.disk.image.files	Extract disk image files during ingest. Please see the **StandardWorkflows** document for more information on Disk Image Extraction.
identify.duplicates.during.ingest	Identify duplicates during Ingest by throwing an overridable error during the Fixity Check step.
minimum.comment.length	Minimum length that comments are allowed to be in workflows.
opex.ingest.delay	The timeout period for the OPEX Ingest workflow awaiting content to ingest, in seconds. Default: 300
report.maximum.count	Maximum records to display when viewing a report
single.du.disk.image.extraction	Extract disk image files into single folder
wayback.render.timeout	Wayback Render timeout in seconds. The default value is 120 seconds.

4.5. Reference Metadata

Reference metadata allows users to link metadata to entities, so that multiple entities can refer to the same metadata, but that metadata can be updated without having to change every entity that refers to it. See the User Guide for more about the user perspective.

If you have the Manager role, the Reference Metadata page will present extra options to allow you to configure the reference metadata tables available for all users on your tenancy.

You can **create a new table** with the button at the top. Tables can't currently be deleted or renamed, or have their security updated once they're created, so make sure you get the name and security descriptor for the table correct. Once you create the new table it will appear in the list of tables.

Each table will also have a **Configure Table Fields** button, which will take you into an editing mode to configure the fields of the table. The main part of this panel is an interactive listing of the fields:

- To add a new field, use the + buttons at the end of each row; this will insert a new field beneath the current one.
- To edit a field, use the pencil button. (A new field will automatically be put into edit mode.) When editing, you can change the properties of the field, although if the field is in use (i.e. there is at least one entry in the table which has a value for it), you can’t change the internal name or type of the field. Use the Save and Cancel buttons to end edit mode for that row.
- To re-order fields, use the up and down arrow buttons on each row.
- To delete a field, use the Delete button. It is only possible to delete a field if no records are using it.
- To change the display field (the field from which the value shown when linking to a record is taken), press the button on the appropriate row.

4.5.1. Connections

Connections allow you to link a reference metadata table to a field within descriptive metadata (as specified in a custom indexer). Use the dropdown and the + button to add a connection. This will result in the properties page within Explorer showing the field you select as a link (in the viewer) or dropdown (in the editor) to the relevant reference metadata record, instead of showing the raw value in the metadata field.

Each metadata field may only be used in one connection, and the connection will apply to all fragments which have that field in the whole tenancy.

Use the Delete button next to an existing connection to remove it.

4.6. Retention Management

Retention policies allow users to protect entities from deletion and when the policy expires delete or review entities.

If you have the Manager role, the Retention Management page will present policies with the options to modify them or create a new policy.

Policies which have entities assigned to them cannot be changed or deleted other than to disable the ability to assign new entities to them.

When creating or editing a policy you have the follow options:

- Name
- Description
- Period (how long to apply the policy)
- From Field (field used for the start of the policy, available fields are xip.created and any date field specified in a custom indexer)
- Prevent Deletion (whether to prevent deletion of an entity while the policy is in effect)
- Action (what action to perform when the policy expires)

note

a policy must prevent deletion or perform some action on expiry (or both).

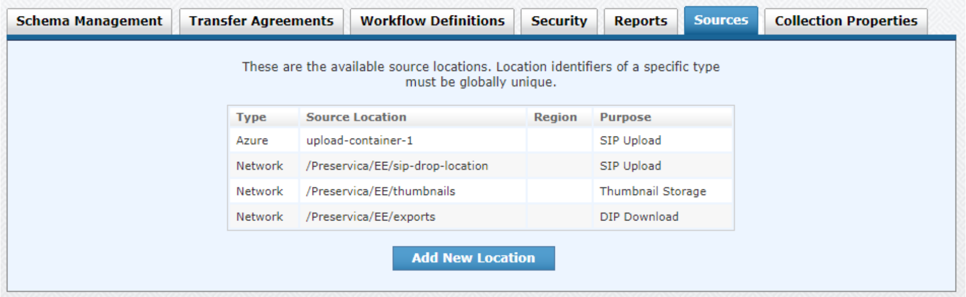
Policies can be assigned to entities through Explorer / Search see the User Guide for more information.

4.7. Sources

The Sources page lets you configure where the system will find or save content which isn’t the primary material that gets stored on storage adapters. Sources are used for:

- SIP Upload: A location into which you can upload ingest packages. These locations are used in ingest workflows. Submission users will need to be able to access this location. PUT, SDB and JobQueue apps require access.
- DIP Download: A location into which the system will save download packages when running an export workflow. Users don't need direct access to this location, as the Download tab of the Access workflow management area will allow a download through the browser. SDB, JobQueue and API apps require access.
- Thumbnail Storage: Where thumbnails are stored. You don't need direct access to this location. Explorer, JobQueue and API apps require access.
- PUT Storage: A location which PUT uses to manage content within its project. PUT will only use the first location defined on the tenancy. You should not directly access this location. Only PUT app requires access.
- PUT Holding: An area where users can upload and manipulate content before bringing it into a project. PUT will only use the first location defined on the tenancy. Users should have access to this location to upload and edit their content. Only PUT app requires access.

Figure 4.11. Sources table



These are the available source locations. Location identifiers of a specific type must be globally unique.

Type	Source Location	Region	Purpose
Azure	upload-container-1		SIP Upload
Network	/Preservica/EE/sip-drop-location		SIP Upload
Network	/Preservica/EE/thumbnails		Thumbnail Storage
Network	/Preservica/EE/exports		DIP Download

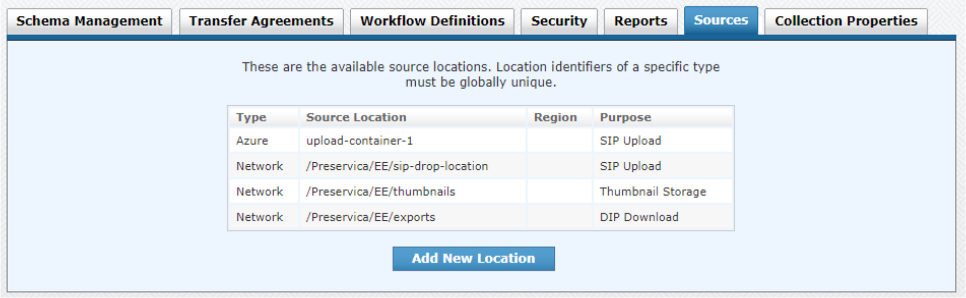
To add a new location, press the Add New Location button and configure it appropriately. The *Type* dropdown will show Network, and also any cloud credentials that have been set up by an administrator (see Section 4.7.1). The source location is the path or bucket/container name. If you are setting up an S3 location, make sure the region is set appropriately for the location of your Preservica server.

You can't edit or remove a location after adding it. Make sure the configuration is correct before saving!

4.7.1. Credentials

The credentials section is only shown to administrators (i.e. with the `SDB_ADMIN_USER` role). It allows you to set up cloud credentials which are used by locations. This section shows, and allows credentials to be set up, for *all tenancies*.

Figure 4.12. Credentials table



These are the available source locations. Location identifiers of a specific type must be globally unique.

Type	Source Location	Region	Purpose
Azure	upload-container-1		SIP Upload
Network	/Preservica/EE/sip-drop-location		SIP Upload
Network	/Preservica/EE/thumbnails		Thumbnail Storage
Network	/Preservica/EE/exports		DIP Download

To add a new credential pair, press the button and fill out the details. Like locations, you can't edit or remove credentials later, so make sure you have got them right.

4.8. Recoverable Deletion

When a workflow is run that deletes original content and metadata the records will enter a cool off period (called the retention period). During this period content and metadata will appear to be deleted and will not be accessible to users. The audit trail will also show the deletion event. The record will only be finally deleted after the retention period has expired and the *Final Deletion Workflow* (Section 4.8.3) is run. This workflow must be set up for this feature to work.

However, it will be possible to restore content before the retention period expires by running the *Restore Deleted Content* workflow described in Section 4.8.4.

By default the retention period is set to 90 days.

4.8.1. Setting the tenancy specific retention period

By default the recoverable deletion retention period is set to 90 days in each tenancy. To set a retention period to another value in days, the tenant specific system property must be set by an administrator. If a tenancy of *EXAMPLE_TENANT* exists the property *delete.retention.period* should be added with the tenancy string set to *EXAMPLE_TENANT* and the retention period set to any positive integer, or zero. A record's retention date is not updated when this parameter is changed; it is set when the workflow is run.

Attempting to set the retention period to below zero days or with an invalid value will set the retention period to the default period of 90 days.

4.8.2. Reporting Recoverable Deletions

The **Records Pending Deletion report** has been added which will allow a system administrator to view basic information about the deleted content, including when it was deleted, the title of the deleted record, the type of deletion and the date when the retention period is over. The two types of deletion are reported as *FULL* (content and metadata), and *CONTENT* (content only, leaving metadata).

4.8.3. Final Deletion

The **Final Deletion workflow** (data.management.final.deletion.rf) must be set up to remove content and metadata which has passed its retention period. Content will be deleted from the disk adapters for each recoverable deletion. For *FULL* recoverable deletions, the entity records and everything associated with them are also deleted, and the item will no longer be reported as a recoverable deletion. Deletions are permanent and immediate.

The audit trail for the entities will also be deleted in the case of a full deletion. Records are created in the deletion history tables to record information about the deletion and the appraisal decision related to it.

4.8.4. Restore Deleted Content

The **Recover Deleted Records workflow** definition (data.management.recover.from.deletion.rf) allows a user to restore all content and metadata still within its retention period. For deletions performed on a folder, only the top level record can be selected; partial restoration is not permitted.

If a *FULL* recoverable deletion is selected for recovery, any parent folders that were separately *FULL* deleted will also be recovered. The same applies when restoring a content deletion.

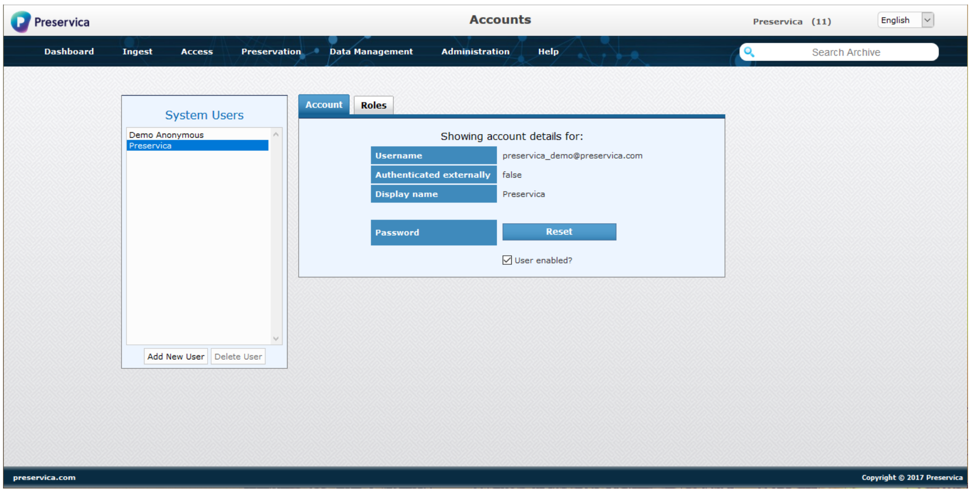
Once a restoration has been performed, the content and metadata can be accessed again. The audit trail will include the deletion event and restoration event.

4.9. Manage Accounts

Depending on how Preservica has been deployed and configured, a **Manage Accounts** option may also be present on the **Administration** menu. This provides access to the User Management module.

The User Management module is an easy-to-use front-end for managing user accounts in the back-end user credentials store (e.g. OpenLDAP). User accounts can be created, modified and deleted from this page by admin users.

Figure 4.13. Manage Accounts Page



Demo Anonymous
Preservica

Showing account details for:

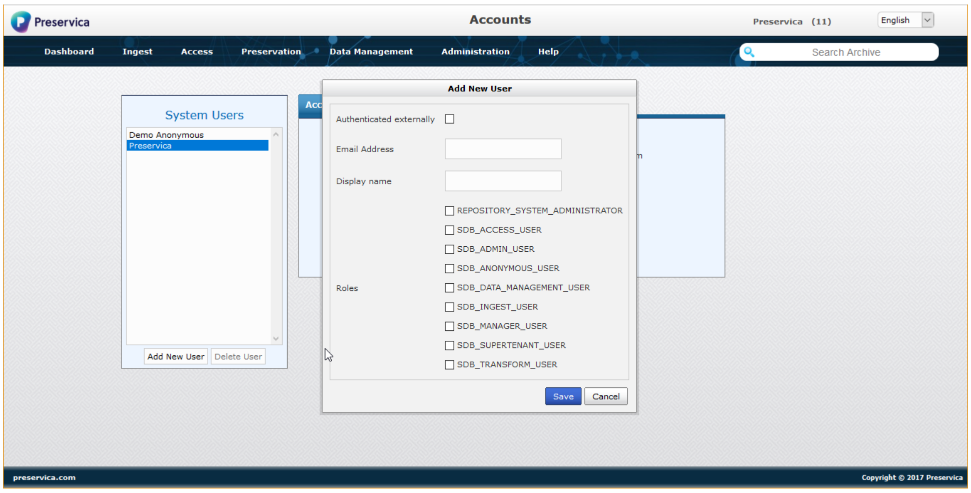
Username	preservica_demo@preservica.com
Authenticated externally	false
Display name	Preservica
Password	Reset

☑ User enabled?

4.9.1. Adding a New User

To add a new Preservica user, simply click on the **Add New User** button under the list of existing users. In the dialog that pops up, enter a valid email address (this also acts as the user name for the new account), a display name, and select which functional and data access roles should be assigned to the user. Then click the **Save** button; the new user is added to the list on the page.

Figure 4.14. Adding a New User



Demo Anonymous
Preservica

Add New User

Authenticated externally □

Email Address

Display name

Roles
- □ REPOSITORY_SYSTEM_ADMINISTRATOR
- □ SDB_ACCESS_USER
- □ SDB_ADMIN_USER
- □ SDB_ANONYMOUS_USER
- □ SDB_DATA_MANAGEMENT_USER
- □ SDB_INGEST_USER
- □ SDB_MANAGER_USER
- □ SDB_SUPERTENANT_USER
- □ SDB_TRANSFORM_USER

Passwords

A strong password is automatically generated and emailed to the user. On their first login to Preservica they are redirected to their account page and recommended to change their password. Passwords must be between 8 - 64 characters in length, and contain at least 3 out of the following:

- uppercase letters
- lowercase letters
- numbers
- symbols

4.9.2. Deleting a User

To delete an existing Preservica user, select their name in the list and click on the **Delete User** button. Click OK on the confirmation dialog; the user account will then be permanently deleted from the system.

4.9.3. Modifying an Existing User’s Password

A user’s password can be reset by selecting the user in the list on the left of the page, and then clicking the password **Reset** button. This will generate a strong password for the user, and send it to them via email. As noted above, the next time the user logs in to Preservica, they will be redirected to their account page and recommended to change their password.

4.9.4. Enabling / Disabling a User Account

Depending on how the back-end user credential store (e.g. OpenLDAP) is configured, a user account may be automatically disabled for a period of time after a number of unsuccessful login attempts. If this is the case, then when that user is selected in the list, the **User enabled?** checkbox will reflect the account status; i.e. if the account has been temporarily disabled, the checkbox will be unticked.

To re-enable a user account, simply tick the **User enabled?** checkbox. (A confirmation message will then be displayed.)

To disable a user account, select the user in the list and then untick the **User enabled?** checkbox. (Again a confirmation message is then displayed.) This will *permanently* disable the account until the **User enabled?** checkbox is ticked once more.

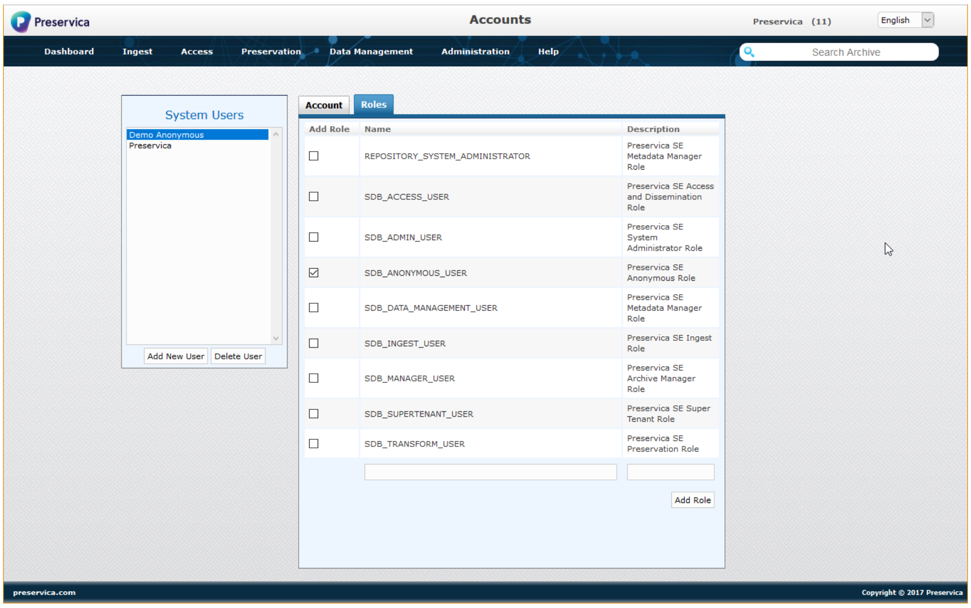
note

Enabling / disabling a user account has no effect on the account password.

4.9.5. Assigning Roles to a User Account

To assign and remove functional and data access roles to/from a user account, select the user in the list on the left of the page, and then click on the **Roles** tab. This will display a list of the currently defined roles, with those currently assigned to the user ticked.

Figure 4.15. Assigning Roles to a User



Add Role	Name	Description
□	REPOSITORY_SYSTEM_ADMINISTRATOR	Preservica SE Metadata Manager Role
□	SDB_ACCESS_USER	Preservica SE Access and Dissemination Role
□	SDB_ADMIN_USER	Preservica SE System Administrator Role
☑	SDB_ANONYMOUS_USER	Preservica SE Anonymous Role
□	SDB_DATA_MANAGEMENT_USER	Preservica SE Metadata Manager Role
□	SDB_INGEST_USER	Preservica SE Ingest Role
□	SDB_MANAGER_USER	Preservica SE Archive Manager Role
□	SDB_SUPERTENANT_USER	Preservica SE Super Tenant Role
□	SDB_TRANSFORM_USER	Preservica SE Preservation Role

To modify the roles assigned to the user, simply tick / untick the roles as required.

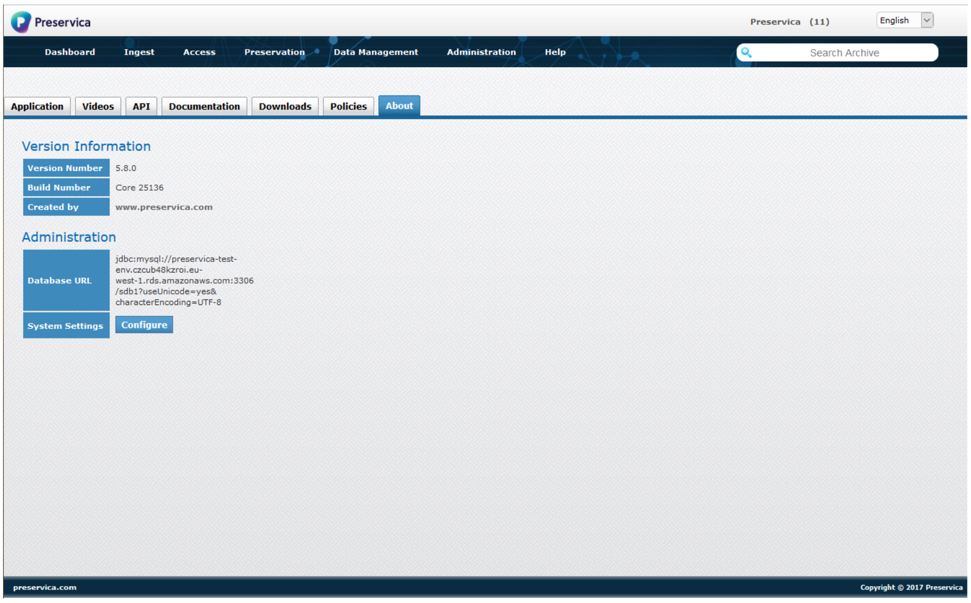
A new role can be added to the system by entering a name and description for the new role in the text boxes at the bottom of the list, and then clicking on the **Add Role** button.

4.10. About

The **About** tab is found under the **Help** menu and displays the Preservica **Version Number** and **Build Number(s)** (see Figure 4.16). This information may be requested by support personnel when investigating a reported problem.

If the user has the administration permissions (`ROLE_SDB_ADMIN_USER`), then the Database URL is shown for information and the user can access the system configuration pages (see Chapter 5) by pressing the **Configure** button.

Figure 4.16. About Page



Version Information

Version Number	5.8.0
Build Number	Core 25136
Created by	www.preservica.com

Administration

Database URL	jdbc:mysql://preservica-test-env.czcub48kzroi.eu-west-1.rds.amazonaws.com:3306/sdb1?useUnicode=yes&characterEncoding=UTF-8
System Settings	Configure

Chapter 5. Configure

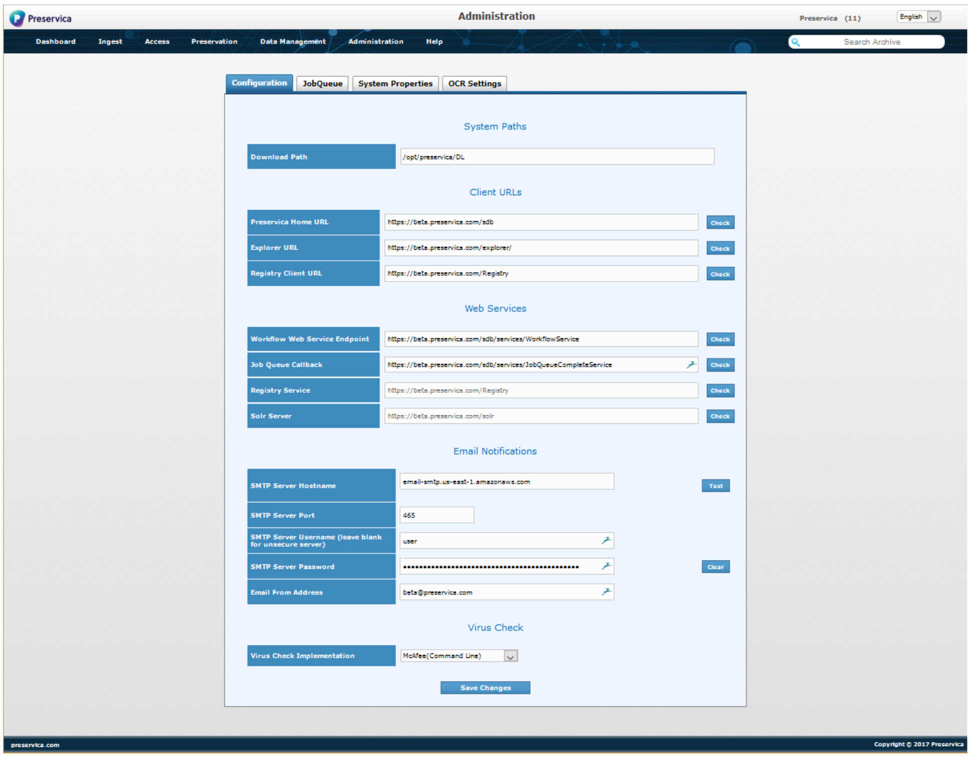
The configuration pages are only available to administrators (i.e. with the `ROLE_SDB_ADMIN_USER` role), not managers, and configure system wide settings.

5.1. Administration Page - Configuration Tab

The configuration tab (see Figure 5.1) allows an administrator to set the basic parameters that control the operation of the Preservica System.

Any changes made should be saved using the **Save Changes** button at the bottom of the form.

Figure 5.1. Preservica Configuration Page



System Paths

Download Path	/opt/preservica/DL

Client URLs

Preservica Home URL	https://beta.preservica.com/sdb
Explorer URL	https://beta.preservica.com/explorer/
Registry Client URL	https://beta.preservica.com/Registry

Web Services

Workflow Web Service Endpoint	https://beta.preservica.com/sdb/services/WorkflowService
Job Queue Callback	https://beta.preservica.com/sdb/services/JobQueueCompleteService
Registry Service	https://beta.preservica.com/Registry
Solr Server	https://beta.preservica.com/solr

Email Notifications

SMTP Server Hostname	email-smtp.us-east-1.amazonaws.com
SMTP Server Port	465
SMTP Server Username (leave blank for unsecure server)	user
SMTP Server Password	••
Email From Address	beta@preservica.com

Virus Check

Virus Check Implementation	McAfee(Command Line)

5.1.1. System Paths

The System Paths section of the Configuration tab (see Figure 5.1) is where you can set all the paths used by Preservica.

- Download Path is the location that will be used to store generated DIPs, ready for download. This location is only used for pre-6.0 DIPs; any new DIPs will be written to the Source (see Section 4.7) specified in the export workflow.

5.1.2. Client URLs

The Client URLs section of the Configuration tab (see Figure 5.1) stores the URLs of the various applications in the Preservica system, from the client's point of view. (These are not necessarily the same as the URLs from the server's viewpoint if a reverse proxy is being used.)

- Preservica Home URL is the URL at which users can reach Preservica. (This is used in email links sent from Preservica.)

- Explorer URL is the URL at which users can reach the Explorer application for browsing the contents of the archive.

- Registry Client URL is the URL for the technical registry (LDR).

5.1.3. Web Services

The Web Services section of the Configuration tab (see Figure 5.1) stores the URLs of various web services and servers used within the Preservica system:

- Workflow Web Service Endpoint is the URL for starting Preservica workflows.
- Job Queue Callback is the location of the web service listener on which Preservica listens for returning Job Queue responses.
- Registry Service is the URL for the instance of the Registry application (LDR) used by Preservica. This field displays the setting in the shared local.properties file and is not editable.
- Solr Server is the URL for the Solr search server used to index and search the archive’s contents. Again this field displays the setting in the local.properties file and is not editable.

Pressing the **Check** button beside a URL tests the *saved* URL for that server or Web Service. So, to check an updated URL, you first need to save your changes by pressing the **Save Changes** button at the bottom of the page. If the URL being tested is correct the browser will open the appropriate webpage or wsdl in a new browser window or tab. For the Workflow Web Service Endpoint and the Job Queue Callback, the WSDL for the web service is opened. For the Registry Service, the registry home page is opened, while for the Solr server, the Solr admin page is opened. Note that some types of external single sign on using a reverse proxy, or security settings in the Apache configuration, will prevent this check from working. In these cases, if it is necessary to undertake the check, it should be done from a web browser running on the Preservica workflow server.

5.1.4. Email Notifications

The email notifications section of the Configuration tab (see Figure 5.1) contains the configuration details which Preservica uses to send emails to users. The SMTP server username and password options are only required if connecting to a secure mail server. To clear an entered password, press the **Clear** button to the right of the field. Pressing the **Test** button by the SMTP Server Hostname field will test the email settings by sending a test email to the user you are logged in as.

If problems are encountered sending e-mails from Preservica, then the log files associated with any firewall, anti-virus, or domain policy systems should be checked as these systems can prevent the Preservica system from sending e-mails.

5.1.5. Virus Check

Preservica supports the use of Clam AV to scan content for viruses during ingest. See the Linux or Windows installation guide for more information.

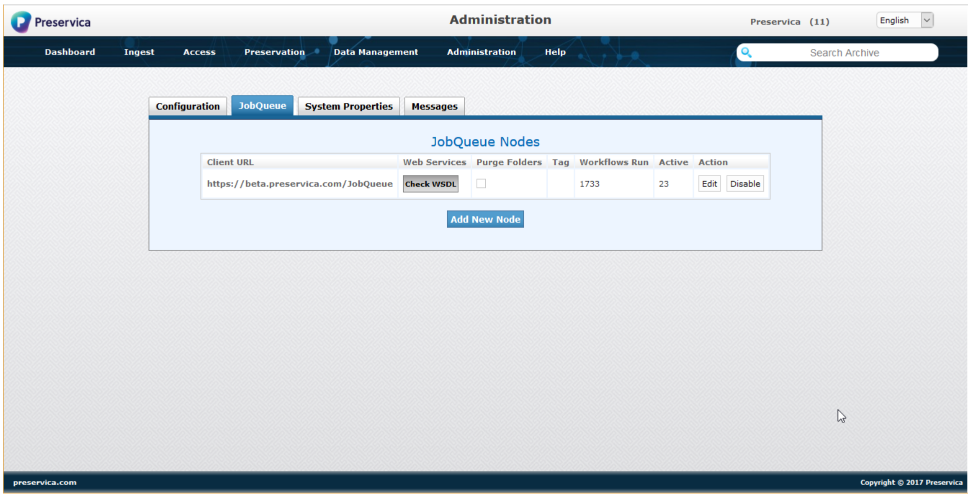
5.2. Administration Page - JobQueue Tab

The JobQueue tab (see Figure 5.2) allows an administrator to add JobQueue nodes to the Preservica system. At least one JobQueue node must exist and be available. Additional JobQueue nodes can be added where it is necessary to distribute the workload across multiple servers for performance reasons. Each workflow is pinned to a specific node and all its constituent workflow steps will be executed on the same JobQueue node.

If multiple JobQueue nodes are present, then by default the Preservica workflow system will apply a simple "round robin" approach to distributing workflows to individual JobQueue nodes. This default approach can be overridden by developing a custom scheduling module and defining this within Preservica’s System Properties (see Section 5.3). For more information on load balancing, see Section 5.2.1.

Preservica also provides an alternative scheduling approach where a **Tag** can be applied to specific workflows definitions or contexts and any workflows instances started will be directed to the Job Queue node with the matching tag.

Figure 5.2. List of Job Queue Nodes



Client URL	Web Services	Purge Folders	Tag	Workflows Run	Active	Action
https://beta.preservica.com/JobQueue	Check WSDL	☐		1733	23	Edit Disable

The list of Job Queue nodes (see Figure 5.2) displays the status of the current Job Queue nodes; this includes the Client URL, the total number of workflows run on the node and the number of workflows currently active on the node.

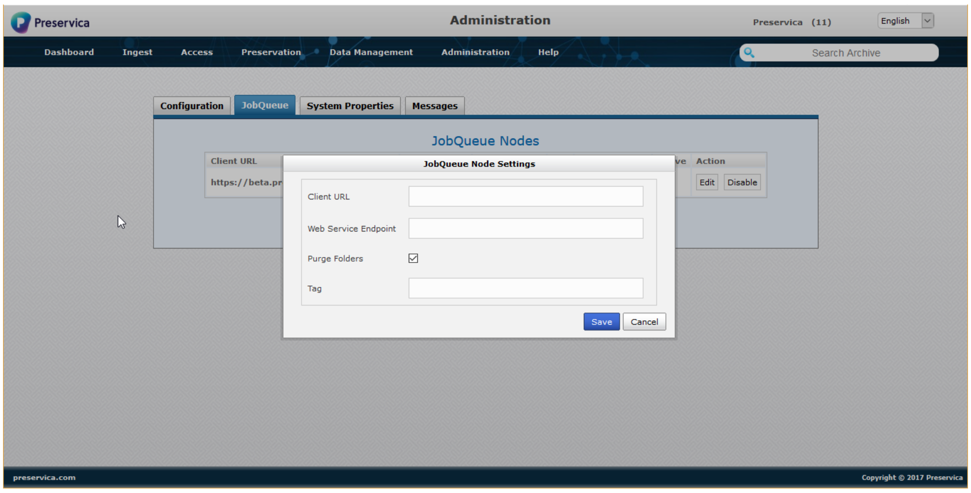
The validity of a node's URL can be checked by pressing the **Check WSDL** button, which, if the URL is correct, will open the Job Queue's WSDL in a new browser window (or tab). Note that some types of external single sign on using a reverse proxy will prevent this check working. In these case if a necessary to undertake the check it should be done from a web browser run on the Preservica workflow server.

Nodes can be disabled or enabled as needed by pressing the appropriate button, but at least one Job Queue node must be enabled to allow workflows to be run. Nodes can be removed (by pressing the **Remove** button) *only* if no workflows have been run on them, otherwise they can only be disabled to prevent future workflows from being run on them.

Clicking on the entry in the Client URL field for a node will open the Job Queue management User Interface for it. This interface allows you to view the configuration of the Job Queue node and see what jobs (i.e. workflow steps) are currently running and have been run on the node.

The **Purge Folders** setting and the priority **Tag** for a node can be edited by pressing the **Edit** button for the node. This brings up the same pop-up window as is used for adding a new node (see Figure 5.3), except that the Client URL and Web Service Endpoint text boxes are disabled).

Figure 5.3. Add Job Queue Node



JobQueue Node Settings

Client URL

Web Service Endpoint

Purge Folders ☑

Tag

Use the **Add New Node** button to add a new Job Queue node into the Preservica system, which will bring up a pop-up window (see Figure 5.3) into which you can enter all the information required to set up a new Job Queue node:

- Client URL is the URL for the Job Queue application.
- Web Service Endpoint is the location of the web service listener on which the Job Queue application listens for instructions. This will default to the standard location based on the Client URL but can be amended if necessary.
- If the Purge Folders checkbox is selected, then Job Queue work folders (located under the **Job Queue Working Area** system path) will be deleted once each job has been completed. Preservica should normally operate with the purging of Job Queue Folders enabled.
- An optional tag for load balancing.

Once created, new Job Queue nodes can be managed in the same way as existing nodes.

5.2.1. Load Balancing Strategies

Where more than one JobQueue node is available to the system, Preservica has to determine how workflows should be assigned to a node for processing. By default, a “Round-Robin” load-balancing strategy is used, ensuring that workflows are evenly spread across all available nodes. In addition to this, three other load-balancing strategies are available for use.

Large SIP

This forces all large SIPs (i.e. those with large numbers of entities)[1] to be processed on a designated sub group of JobQueue nodes, allowing these servers to be specified and configured specifically to deal with these SIPs. For small SIPs, and non-ingest workflows, the default Round-Robin strategy is still used.

To use this strategy, the “job.queue.load.balancer.bean” system property (see Section 5.3) should be set to *largeSIPJobQueueLoadBalancer*. The default settings for this strategy are that designated nodes should be tagged “large.sip.node” and SIPs with metadata files bigger than 100MiB will be considered large. These defaults can be over-ridden in the local.properties configuration file by setting the following properties:

- **large.sip.node.tag** - The new tag string.
- **large.sip.threshold** - The size of the metadata file (in bytes) above which a SIP is considered large.

This load balancer will only override the default Round-Robin strategy if the size of the metadata file can be assessed prior to the selection of JobQueue node, this means that:

1. The workflow must be an ingest workflow launched from a file watcher where the SIP is selected automatically.
2. The SIP must be staged on a conventional file system (e.g. not in an S3 bucket, etc.).
3. The SIP must have a standard Preservica layout, such that the metadata file is addressable by the workflow server at **SIP_LOCATION/SIP_ID/metadata.xml**.

Priority Value

This strategy allows you to tag JobQueue nodes with custom “priorities”, creating distinct sub-groups. Workflow definitions can be tagged to match, meaning that workflows based on those definitions will run only on JobQueue nodes from that sub-group; where the sub-group contains more than one node, the selection is made on a round-robin basis from the sub-group.

[1] In fact, it is the size of the XIP metadata file which is checked, not its contents

If no JobQueue nodes match the workflow’s tag, then the workflow will run on the sub-group of untagged JobQueue nodes (falling back to the default round-robin if all nodes are tagged).

Workflow definitions that are not tagged will run on the sub-group of JobQueue nodes that are also untagged. If all JobQueue nodes are tagged, untagged workflows will fallback to the default round-robin strategy (and thus will run on one of the tagged nodes).

To use this strategy, the “job.queue.load.balancer.bean” system property (see Section 5.3) should be set to *priorityValueJobQueueLoadBalancer*. Any workflow definitions that should take advantage of the JobQueue tags should contain the following variable:

```
<variable name="constPriorityValue" >
  <type name="org.drools.process.core.datatype.impl.type.StringDataType"/>
  <value>MyTag</value>
</variable>
```

Where “MyTag” is the string used to tag the JobQueue node.

Multi-Site

This strategy allows you to define two pools of JobQueue servers and is intended for use where the two pools are in geographically distinct locations and content in each location should be processed locally. In this case, each JobQueue node should be tagged with a location, and workflow contexts should contain the location in which they should be run as part of their name or description. This strategy also ensures that regular integrity checking gets run in the correct location.

To use this strategy, the “job.queue.load.balancer.bean” system property (see Section 5.3) should be set to *multiSiteJobQueueLoadBalancer*. The default settings for this strategy are that the locations should be called “siteA” and “siteB”. These defaults can be over-ridden in the local.properties configuration file by setting the following properties:

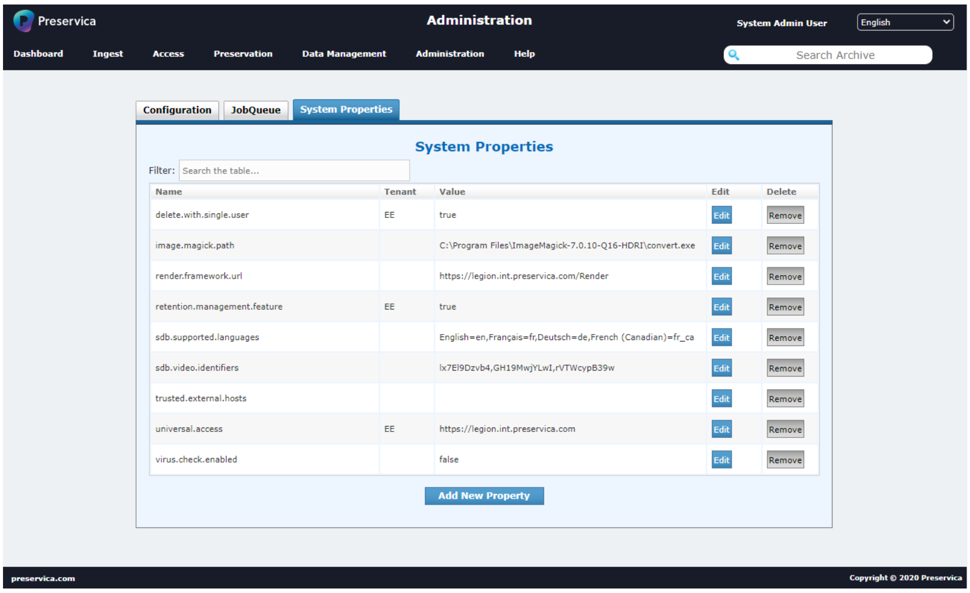
- **siteA.tag** - The location of site A.
- **siteB.tag** - The location of site B.

If no JobQueue nodes are tagged, this strategy falls back to the default round-robin. If a particular workflow is not tagged as either site, this strategy will fall back to the default round-robin selection.

5.3. Administration Page - System Properties Tab

The System Properties tab provides a means to configure the system behaviour and specify the location of tools external to Preservica. In previous versions of Preservica/SDB, a number of the system properties were set in a configuration file, but these are now all set from the System Properties tab. The System Properties tab allows an administrator to create and remove system properties in a controlled manner without needing to restart Preservica for the changes to take effect. The System Properties tab provides a list of the properties that have been set and their associated values (see Figure 5.4). Existing System Properties can be edited by pressing the appropriate **Edit** button and removed by pressing the appropriate **Remove** button.

Figure 5.4. List of System Properties



System Properties

Filter: Search the table...

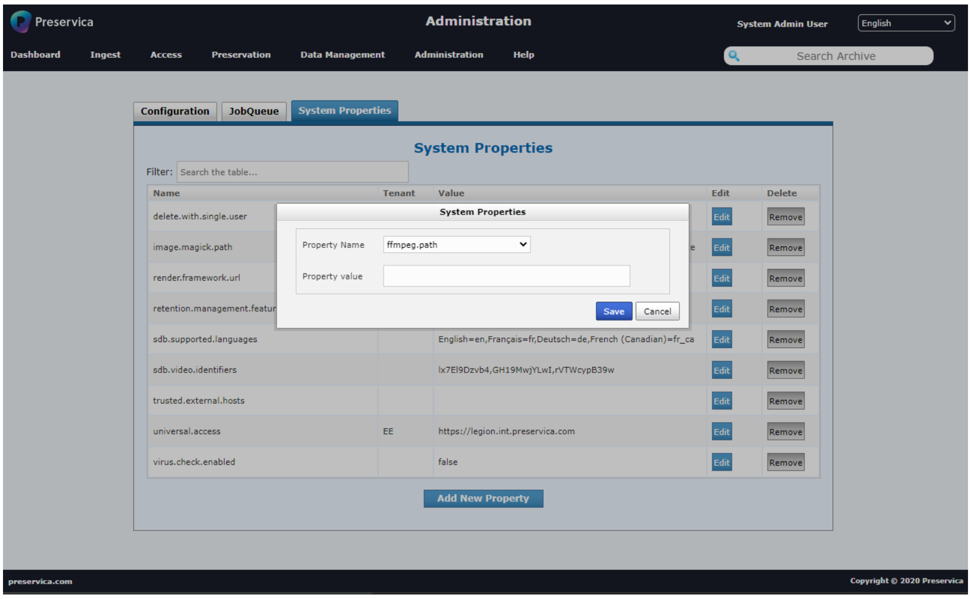
Name	Tenant	Value	Edit	Delete
delete.with.single.user	EE	true	Edit	Remove
image.magick.path		C:\Program Files\ImageMagick-7.0.10-Q16-HDRI\convert.exe	Edit	Remove
render.framework.url		https://legion.int.preservica.com/Render	Edit	Remove
retention.management.feature	EE	true	Edit	Remove
sdb.supported.languages		English=en,Français=fr,Deutsch=de,French (Canadian)=fr_ca	Edit	Remove
sdb.video.identifiers		lx7El9Dzvb4,GH19MwjYLwI,rVTWcypB39w	Edit	Remove
trusted.external.hosts			Edit	Remove
universal.access	EE	https://legion.int.preservica.com	Edit	Remove
virus.check.enabled		false	Edit	Remove

Additional System Properties can be set by using the **Add System Property** button. Pressing this displays the **Add System Property** dialog box (see Figure 5.5). The required property can be selected from the **Property Name** drop-down list and the required value entered into the **Property Value** field. The list of Property Names does not include properties for which a value has already been set to prevent setting duplicate properties. When entering file and path names please remember that these may be case sensitive depending on the operating system in use. The use of the forward slash ‘/’ as a directory delimiter is recommended, rather than the backward slash ‘\’, since the forward slash will be interpreted correctly on all supported operating systems.

For properties which can have custom settings on a tenant-by-tenant basis, a **Tenant** field will become visible.

Within the list of configurable System Properties are a number of properties whose keys are suffixed by ".feature". These properties can be used to enable access to particular features of Preservica for the user. These features are by default set to be false, and so the absence of an entry for a feature in System Properties will render that feature disabled - until an entry is added, with the value of "true".

Figure 5.5. Add System Property



System Properties

Filter: Search the table...

Name	Tenant	Value	Edit	Delete
delete.with.single.user			Edit	Remove
image.magick.path			Edit	Remove
render.framework.url			Edit	Remove
retention.management.featur			Edit	Remove
sdb.supported.languages		English=en,Français=fr,Deutsch=de,French (Canadian)=fr_ca	Edit	Remove
sdb.video.identifiers		lx7El9Dzvb4,GH19MwjYLwI,rVTWcypB39w	Edit	Remove
trusted.external.hosts			Edit	Remove
universal.access	EE	https://legion.int.preservica.com	Edit	Remove
virus.check.enabled		false	Edit	Remove

System Properties

Property Name: ffmpeg.path

Property value:

Save Cancel

Table 5.1 describes all the available system properties.

Table 5.1. System Properties

System Parameter	Notes
auto.recharacterise.feature	Set to *true* to enable the Auto Re-Characterisation feature. This can be set globally on a system or can be enabled for specific tenants.
auto.workflow.trigger.prefix	An option to specify a trigger file prefix for the S3 Bucket Event monitor.
batch.size.limit	Maximum size of each batch in workflows that download content (to the server) and use batching.
clam.av.path	Path to the ClamAV antivirus engine.
co.order.feature	Set to *true* to enable ordering of Content Objects.
dcraw.path	Full path (drive (on Microsoft Windows OSes), directories and filename) to the dcraw executable: "dcraw.exe".
delete.allow.forced.final	Enable or disable forced final deletion of one or more entities in the final deletion workflow before the retention period has elapsed. This functionality can only be triggered through the API. Disabled by default.
delete.retention.period	Cooling off period in days for recoverable-deleted content. See section Section 4.8 for more details.
delete.with.single.user	Enable or disable single user deletion in the standard deletion workflow by automatically approving the request. Disabled by default.
droid6.max.bytes.to.scan	The maximum number of bytes that Droid6 will scan when identifying files. If set to a negative integer, then the entire file will be scanned.
embed.jhove.xml.metadata	If set, then following characterisation JHOVE metadata will be inserted into the XIP metadata for each file.
enable.exi.encoding	If set, then metadata fragments will be stored in the database using Efficient XML Interchange encoding. This will reduce the amount of storage space used significantly. Only applies if metadata is stored in a database.

System Parameter	Notes
entity.task.batch.size	Maximum size of each batch of entity tasks.
exif.tool.path	Full path (drive (on Microsoft Windows OSes), directories and filename) to the Exif Tool executable.
explorer.number.records	The number of items to display per page in the Explorer Browse view (and Assign Thumbnails dialog). If not set, Explorer will default to 20 items per page.
explorer.record.label	Display label for explorer records. Should be *code* or *title*. If not set, titles will be displayed.
explorer.search.page.limit	The maximum number of results that should be displayed in the Explorer Search view. This number will be used to generate a range of available page sizes that will appear in the results per page drop-down list in the Explorer Search view. If no value is set, then Explorer will default to 10, 20, 30, 40 as the available page-sizes. There is a hard coded limit of 250 which cannot be exceeded.
explorer.transitions.duration	The duration in milliseconds of transition animations in Explorer. Set to 0 to disable transition animations.
export.feature	Set to *true* to enable the Export feature.
export.limit	The maximum number of assets that can be included in a single export DIP (tenant specific).
export.size.limit	Maximum size (MB) for a package produced by exporting content (1024MB = 1GB).
external.auth.shared.key	Shared key used to verify external authentication requests, e.g. when using UA with SAML.
ffmpeg.acodec	FFmpeg audio codec preset value
ffmpeg.path	Full path (drive (on Microsoft Windows OSes), directories and filename) to the FFmpeg executable "ffmpeg.exe". Previously, this was held in the local.properties file.
ffmpeg.timeout	Maximum time (in seconds) to allow the FFmpeg or HandBrake executables to run for.
ffmpeg.vcodec	FFmpeg video codec preset value
fiwalk.timeout	Maximum time (in seconds) to allow the Fiwalk tool to run for.
fiwalk.tool.path	Full path (drive (on Microsoft Windows OSes), directories and filename) to the FiWwalk executable "fiwalk.exe".
fulltext.index.blacklist	Comma-separated list of format puids for which full text indexing will not be done. If set, then Preservica will attempt to index all other formats.
fulltext.index.whitelist	Comma-separated list of format puids for which full text indexing will be done. If set, then Preservica will NOT index any other formats.
ghostscript.path	Full path (drive (on Microsoft Windows OSes), directories and filename) to the GhostScript executable
handbrake.path	Full path (drive (on Microsoft Windows OSes), directories and filename) to the HandBrake executable "HandBrakeCLI.exe"
handbrake.timeout	Maximum time (in seconds) to allow the Handbrake tool to run for.
icat.timeout	Maximum time (in seconds) to allow the Icat tool to run for.
icat.tool.path	Full path (drive (on Microsoft Windows OSes), directories and filename) to the FiWwalk executable "icat.exe".

System Parameter	Notes
image.magick.path	Full path (drive (on Microsoft Windows OSes), directories and filename) to the Image Magick executable: "convert.exe".
image.migration.threshold	Value between 1 and 0 to specify the tolerance applied to image histogram comparisons on image migration. A value of 1 will result in the acceptance of all migrations. An initial value of 0.1 to 0.05 is recommended. In addition, the default action for the SDB_IMHC_01 error for migration workflows should be set to “Halt Workflow”.
image.magick.migration.timeout	Maximum time (in seconds) to allow the Image Magick tool to run for when doing migration.
image.magick.rendering.timeout	Maximum time (in seconds) to allow the Image Magick tool to run for when doing rendering.
ingest.asset.duplicate.check	When ingesting an OPEX, how to manage duplicate assets (e.g. by source ref). Can be *global* (the default): files in the source won’t be ingested if they are already anywhere in Preservica; *local*: files in the source won’t be ingested if they are already in the same folder; or *none*: files in the source will always be re-ingested.
ingest.folder.duplicate.check	When ingesting an OPEX, how folder matching should be applied: *SourceIDFirst* (default) will try to match folders globally by the source ID, if one is provided; *TitleOnly* will ignore any source ID and always attempt to match locally based on title.
integrity.manager.email	The (tenant specific) email address for a tenant manager to receive integrity check emails. If emails are configured to be sent in the storage integrity configuration, they will be sent here.
integrity.operations.email	The email address for system operations staff to receive integrity check emails. Emails will always be sent here about failed integrity checks, on any tenancy.
jhove.buffer.size	Overrides default JHOVE buffer size to optimise characterisation performance. It is not necessary to set this parameter in most circumstances, as the default value is appropriate.
jhove.timeout	Maximum time (in seconds) to allow the JHOVE executables to run for.
job.queue.load.balancer.bean	Name of a Spring bean used to provide a custom Job Queue load balancer strategy. If not set, then the default ‘round robin’ approach will be used.
libreoffice.path	Full path (drive (on Microsoft Windows OSes), directories and filename) of the Libre Office executables.
libreoffice.timeout	Maximum time (in seconds) to allow the Libre Office tool to run for migrations and rendering. The default is 120 seconds.
lotus.notes.rmi.export.folder	Full path (drive (on Microsoft Windows OSes), directories and filename) to Lotus Notes RMI export folder.
lotus.notes.rmi.ssl	When set, then the RMI connection to the Lotus Notes exporter process will use SSL.
lotus.notes.rmi.url	URL for Lotus Notes RMI connection.
media.info.path	Full path (drive (on Microsoft Windows OSes), directories and filename) to the Media Info executable: “MediaInfo.exe” (“MediaInfo” on linux).
monitor.feature	Set to *true* to enable the Process Monitoring application (/monitor)

System Parameter	Notes
multipart.download.block.size.bytes	The block size to use for stream copies within multipart downloads (default: 1MB). Doesn't apply if using the S3 Transfer Manager.
multipart.download.part.size.bytes	The part size to use for multipart downloads (default: 128MB). Doesn't apply if using the S3 Transfer Manager.
ocr.feature	Set to *true* to enable OCR.
opex.ingest.delay	Minimum time (in seconds) for items to be in source directory before being eligible for ingest via Bulk Uploader. Default 300 (5 minutes).
opex.ingest.max.bytes	Threshold for splitting OPEX contents into a batch based on file size (in bytes). Default 200GB.
password.operations.email	The email address users are directed to if they receive unwanted password reset emails.
password.reset.link.url	The url for the password reset page.
physical.asset.feature	Set to *true* to enable the creation of physical assets from Explorer and Entity API.
put.url	Base URL for PUT
raw2nexus.path	Full path (drive (on Microsoft Windows OSes), directories and filename) to the raw2nexus program, which is a specialist scientific format conversion tool.
render.framework.url	The URL to the render framework application, for example, https://us.preservica.com/Render. The render server needs to be accessible to browsers.
replace.feature	Set to *true* to enable Content Object replacement.
report.maximum.count	Number of records to display when viewing a Report. If this value is an invalid number or not set, then the default will be set to 10000. NOTE: A tenant specific system property with the same name may also be set. If this is the case, the tenant specific system property will override the system wide system property.
reports.feature	Set to *true* to enable the generation of reports.
rest.entry.formatter.prefix	Name of a Spring bean used to provide custom formatting of the Atom entries in the REST API response. Only applies to the legacy v5 /collections and /deliverableUnits endpoints.
retention.expiry.check.time	Time to check for expired retention assignments. Default: 02:00
retention.expiry.daily.limit	Maximum number of expired retention assignments to process per policy per day. Default: unlimited
retention.management.feature	Set to *true* to enable retention management.
s3.download.approach	How to perform large downloads on S3. Should be TransferManager or PreservicaMultipart. Defaults to PreservicaMultipart
sdb.supported.languages	Comma delimited list of supported languages. Each list item should be in the format: displayName=twoLetterCode. E.g: English=en,Français=fr,Deutsch=de
sdb.video.identifiers	YouTube identifiers of SDB help videos.
search.export.limit	Maximum number of results to allow in a CSV export.
security.management.feature	Set to *true* to enable Security Management. This can be performed via the user interface and the API.
storage.data.collection.time	Time (server time) at which storage data collection tasks will be scheduled. Default: 06:30

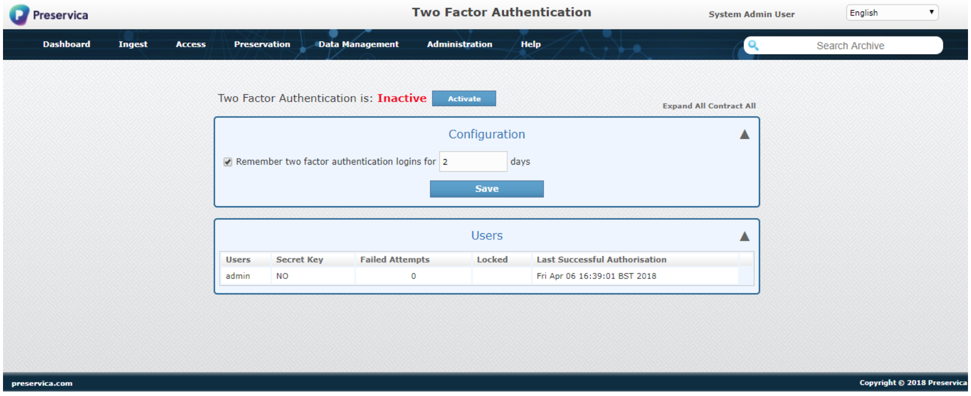
System Parameter	Notes
storage.total.max.quota	Maximum quota (MB) for total storage usage within a tenancy.
tesseract.path	Full path to the tesseract executables (for OCR)
trusted.external.hosts	Comma-separated list of external hosts which are allowed to bypass CSRF filters.
universal.access	Base URL for universal access (tenant specific)
uploaded.object.ingest.delay	Minimum time (in seconds) for objects to be in a network upload source before triggering automatic ingest.
uploadwizard.name	The name of the Upload Wizard installer in the downloads folder.
user.management.feature	Set to *true* to enable the usage of user management API (admin/users).
use.gov.cloud	If set to "true", then the Gov Cloud endpoint will be used when connecting to AWS. ////
use.strict.warc.migration	Forces WARC migration to use conversion records in line with the ISO standard. The ISO standard defines a series of record types such as request records, response records, metadata records and conversion records. Conversion records are required to indicate when an object has been converted from one format to another. The result of a Preservica migration is a second manifestation with the original WARC file, and a second WARC file which only contains conversion records. The idea is that “renderers”, such as the Wayback Machine, are able to determine that the object requested is associated with a conversion record, and will return that instead. However, at present renderers are not able to do this. Therefore, by default, Preservica performs the migration, then modifies the metadata for that record, putting in a http-302 response, which redirects to the new object. The drawback with this approach is the idea that the WARC file describes the http-request and response objects so you get an exact playback, i.e. “when I request url test.com/logo.png, this is the exact response I received”. Effectively this behaviour has now been altered, and in doing so “corrupted” the metadata (of course we still have the original WARC file so the original data is not lost…). However, the Wayback Machine and modern web browsers know how to interpret the http-302 redirect response, so the corrupted version gives the “correct” behaviour upon rendering. Therefore, in Preservica the default is the ISO standard non-compliant migration, since in practice its results can be rendered, whereas the ISO standard compliant migration results cannot. ////
verapdf.path	Path to the Vera PDF CLI tool
verapdf.timeout	Maximum time (in seconds) to allow the Vera PDF CLI tool to run for
virus.check.enabled	Enable/Disable virus check on ingest.
walkme.url	URL for a WalkMe javascript profile which adds WalkMe assistance to Explorer. Either use the full URL including https://cdn.walkme.com/users/ or the path after https://cdn.walkme.com/users/.
wayback.warc.cache	Maximum size (in MB) of the cache of \WARC files for rendering.
wkhtmltoimage.path	Full filename (directory and filename) to the Web Toolkit html to image conversion utility used for producing thumbnails of web pages.

Chapter 6. Two Factor Authentication

Preservica provides Two Factor Authentication to improve security. Two factor authentication expects users to install an application which can generate TOTP tokens as described in RFC 6238, for example the Google Authenticator app for mobile devices.

Two Factor Authentication can be activated or deactivated by users who are administrators or tenancy managers by clicking on the Activate/Deactivate button found on the Two Factor Authentication section.

Figure 6.1. Two Factor Authentication Admin Page



Two Factor Authentication is: Inactive

Configuration

☑ Remember two factor authentication logins for 2 days

Users

Users	Secret Key	Failed Attempts	Locked	Last Successful Authorisation
admin	NO	0		Fri Apr 06 16:39:01 BST 2018

6.1. Configuration

This allows administrators and tenancy managers to turn on and off Two Factor Authentication session time by ticking or unticking the check box. When an administrator or tenancy manager ticks the check box, they can specify the number of days for which Preservica remembers the Two Factor Authentication logins. If the check box is unticked, Two Factor Authentication will have to be performed every time a user logs in to Preservica.

6.2. User Details

This section allows administrators and tenancy managers to view the list of users who have logged into or been recognized by Preservica. Administrators and tenancy managers are able to reset the Two Factor Authentication secret key as well as unlock users who have failed to perform Two Factor Authentication 3 times in a row.

You can also *forget* a user, which will remove them from the records on this page. Note that forgetting a user does not revoke their access; they will still be able to log in to Preservica and set up two factor authentication again. Update user records in your authentication provider (e.g. LDAP), or the Manage Accounts page if using Preservica user management.

6.3. Troubleshooting

Two factor authentication creates a time based token. If none of your users can successfully log in with two factor authentication, check the server time. Servers should generally be synchronised to an NTP time server to avoid this kind of issue.

Chapter 7. Storage

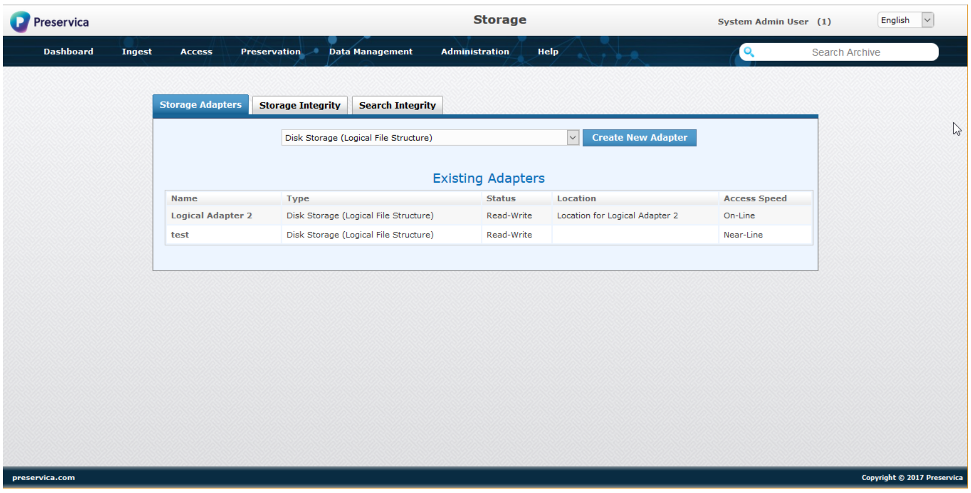
Preservica allows the contents of the archive (i.e. the digital files) to be stored on one or more storage systems, either all content is stored on all storage systems (the default option) or different subsets of the content are stored on different systems according to a set of rules expressed on the storage settings page (see ???). In addition, where there are multiple storage systems, it allows the configuration of which storage system is used for access (normally this would be the fastest access system) and, where appropriate, volume management within a storage system.

Preservica stores content on the storage systems via storage adapters and these can be configured via the administration interface.

7.1. Storage Adapters

The Storage Adapters tab allows you to view, create and edit storage adapters. It shows a list of all storage adapters configured for the system (see Figure 7.1). For each adapter, the adapter's name, type, status, location, and access speed are displayed.

Figure 7.1. Storage Adapter List



Disk Storage (Logical File Structure)

Existing Adapters

Name	Type	Status	Location	Access Speed
Logical Adapter 2	Disk Storage (Logical File Structure)	Read-Write	Location for Logical Adapter 2	On-Line
test	Disk Storage (Logical File Structure)	Read-Write		Near-Line

7.1.1. Adding A Storage Adapter

To add a new storage adapter, select the required adapter type from the drop-down list and then press the **Create New Adapter** button, which will display a form where you can set up a storage adapter; see Figure 7.2 for an example of the form for a disk adapter. See the Storage Adapters guide ([STORE]) for information on the different types of storage adapter available.

The following information must be provided for all storage adapters:

- **Name**: The name must be unique. Two adapters cannot share the same name.
- **Status**: There are four choices for status: Read-Write, Read-Only, Write-Only and Unavailable. If an adapter is unavailable, then no files can be read from or written to it.
- **Access Speed**: The access speed parameter can be used to distinguish adapters for reading purposes. For some adapters only a limited set of access speeds is available. Files are read from the fastest adapter available, i.e. the following order is tried:

  a. On-Line

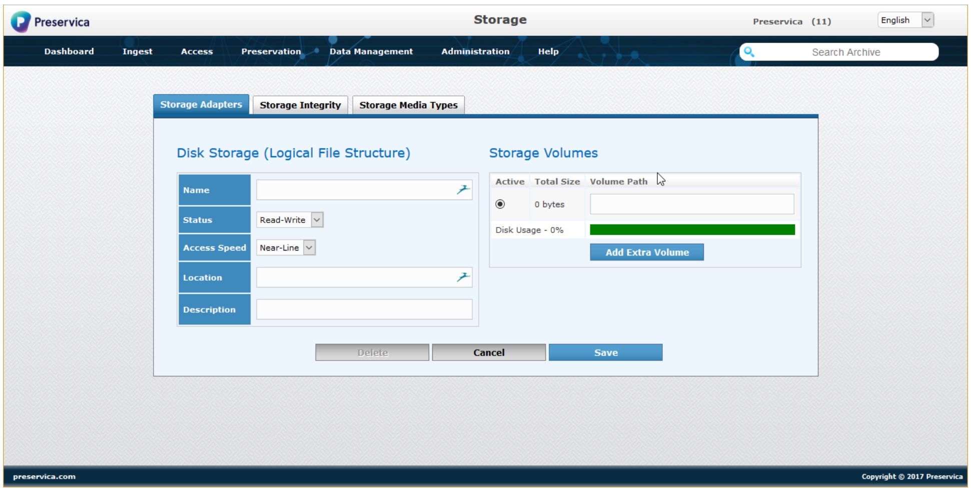
  b. Near-Line

c. Off-Line

- **Location**: A parameter that can be used to describe the location of the adapter in order to help identify it.

Additional, adapter-specific information will be needed to complete the configuration of each adapter (see [STORE] for more details). For example, in Figure 7.2, the right half of the screen displays the disk volumes. The highlighted volume is the one currently in use. The bar underneath displays the disk usage. Additional volumes can be added by pressing the **Add Extra Volume** button. If you add a volume by accident, then leave the path blank and it will not be saved.

Figure 7.2. Add Storage Adapter



Name	
Status	Read-Write
Access Speed	Near-Line
Location	
Description	

Active	Total Size	Volume Path
●	0 bytes	

Disk Usage - 0%

7.1.2. Deleting a Storage Adapter

An adapter that has no archival content saved to it can be deleted from the system using the "Delete" button on the adapter details page. This will remove the adapter, all of its parameters, and any integrity checking configuration for that adapter from the system.

It is **not** possible to delete an existing adapter that already has archival content saved to it in the same manner. This is a safety feature to preserve access to the files in the archive. If an adapter that contained the only copy of files were deleted, then this would destroy the archival integrity. If you no longer wish to use a particular storage system (for reading and writing), then it can be set to Unavailable (rather than deleted).

7.1.3. Updating a Storage Adapter

Extreme caution should be used in amending details of storage adapters and no system activity should be going on when changes are made.

Preservica allows you to modify the parameters of existing storage adapters, as it is plausible that the network address of the adapter may change over time. However, it is possible to make a mistake such that the adapter can no longer access any stored files, which could leave the archive’s integrity damaged. In such circumstances, there may be no other choice than to set the affected adapter’s status to Unavailable, which may result in the permanent loss of access to any files that have only been stored on this affected adapter, and have not been copied elsewhere.

7.2. Storage Integrity

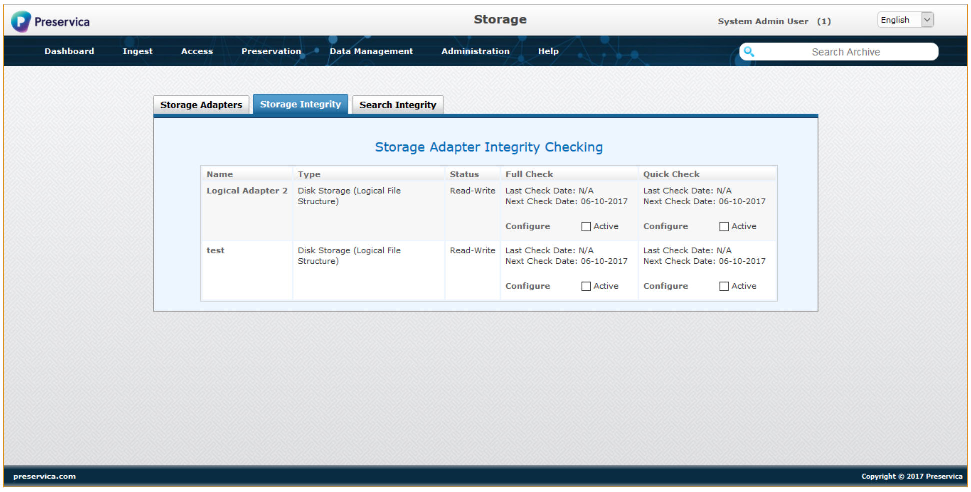
The integrity process works by generating a new fixity checksum for a file and comparing this with the fixity checksum recorded for the file in the metadata store at the point of ingest. If the file has been corrupted or deleted then the integrity process will detect the change in the checksum and will raise an alert. The integrity of all files within a single AIP is checked in a single operation. There is also an option to perform a Quick Check, with the metadata of ingested file sets examined to evaluate whether all files expected files are present and are the correct size.

The **Storage Integrity** tab shows a list of configured storage adapters and allows integrity checking, both the full integrity check and the quick check, to be configured. New storage adapters will have instances of the full integrity check and quick check created automatically when added to the system. The automatically created entries will need to be checked and enabled before use.

For each configured storage adapter, the list (see Figure 7.3) shows the adapter's **Name** and **Type** as set when the adapter was created, to assist in the identification of the correct record to amend. The status of each adapter is also displayed, in terms of the possibility to read and/or write data with these adapters, or simply whether each one is currently unavailable. The list contains two further columns - one each for the full integrity check and the quick check. These columns contain information about the dates of the most recent and next scheduled check.

Adapters for which integrity checking or quick checking is enabled are indicated by ticks in the corresponding **Active** check boxes. A separate check box is present for each check, and for each storage adapter, allowing the user to run either, neither or both checks as per their wishes for each adapter. Performing a particular check on a specific adapter can be enabled or disabled by clicking on the appropriate **Active** check box.

Figure 7.3. Storage Integrity List



Storage Adapter Integrity Checking

Name	Type	Status	Full Check	Quick Check
Logical Adapter 2	Disk Storage (Logical File Structure)	Read-Write	Last Check Date: N/A Next Check Date: 06-10-2017 Configure □ Active	Last Check Date: N/A Next Check Date: 06-10-2017 Configure □ Active
test	Disk Storage (Logical File Structure)	Read-Write	Last Check Date: N/A Next Check Date: 06-10-2017 Configure □ Active	Last Check Date: N/A Next Check Date: 06-10-2017 Configure □ Active

To edit the configuration of either the full or quick check for a particular storage adapter, click on the word **Configure** in the appropriate check's column for the adapter. This will display a pop-up window (see Figure 7.4) where you can amend the Parameters of the particular check:

- **Days between Checks**. Following initial ingest or a successful check, this is the number of days that the system will wait before attempting to check the same content again. As checks are dependent on system capacity this is the minimum number of days between checks as a check on a specific file, once due, could take a number of days to reach the top of the priority list and be checked. The full check has a default setting of 180 days between checks, and the default value for the quick check is 90 days.

- The **Files to check** setting allows the amount of work to be done by the integrity check process to be controlled. Those files which were last checked the longest ago will be checked first.[1] The full check has a default setting of 5,000 files per check, and the default value for the quick check is 50,000 files.

- The full integrity check has the option to repair files which have been discovered to be damaged. If the **Will repair damaged files** check box is ticked, then if the integrity check discovers a file which has

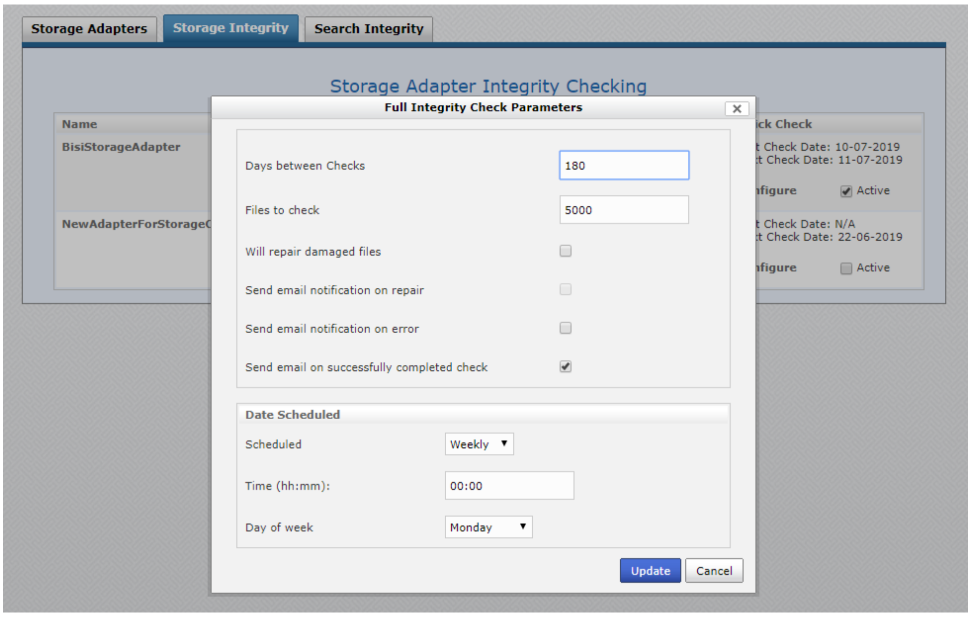
[1] In fact, files are grouped together into arbitrary groups, and it is the groups which are checked, not individual files. This means that the number of files actually checked could vary slightly from the value set. These integrity check groups have no logical meaning.

changed since ingest on one adapter, and there exists another storage adapter on which a non-corrupt copy of the file resides, then Preservica will copy the non-corrupt copy back to the storage adapter for which the full integrity check failed. This feature is not included in the quick check.

- If **Send email notification on repair** is ticked, then when a corrupt or destroyed file is repaired, then an email is sent, to the manager email address specified in the *integrity.manager.email* system property for this tenant, detailing the file that has been repaired and the adapter on which it was repaired. (This option is only present in the Full Integrity Check)
- If the **Send e-mail notification on success** flag is set, then if the check discovers no discrepancies between the expected file set contents and those which are found, the system will send an e-mail to the address specified in the *integrity.manager.email* system property.
- If the **Send e-mail notification on error** flag is set, then if the consistency check identifies an error relating to any of the checked files, the system will send an e-mail to the address specified in the *integrity.manager.email* system property. Note that emails about failures are always sent to the system-wide operator email, whether this checkbox is ticked or not.
- The integrity check process can be **Scheduled** to run on a Daily, Weekly, Monthly or Yearly basis as required. Depending on the frequency option requested, additional information will be requested to define the required schedule.

Once the required changes have been made these can be saved by pressing the **Update** button. Alternatively, to close the pop-up window without saving any changes press the **Cancel** button, or just close the pop-up window.

Figure 7.4. Amend Storage Integrity Settings



Full Integrity Check Parameters

Days between Checks	180
Files to check	5000
Will repair damaged files	□
Send email notification on repair	□
Send email notification on error	□
Send email on successfully completed check	☑

Date Scheduled

Scheduled	Weekly ▾
Time (hh:mm):	00:00
Day of week	Monday ▾

7.3. Search Integrity

The search integrity process is a pair of cross-checks between the Solr index and the metadata held in the metadata store to ensure that they are consistent.

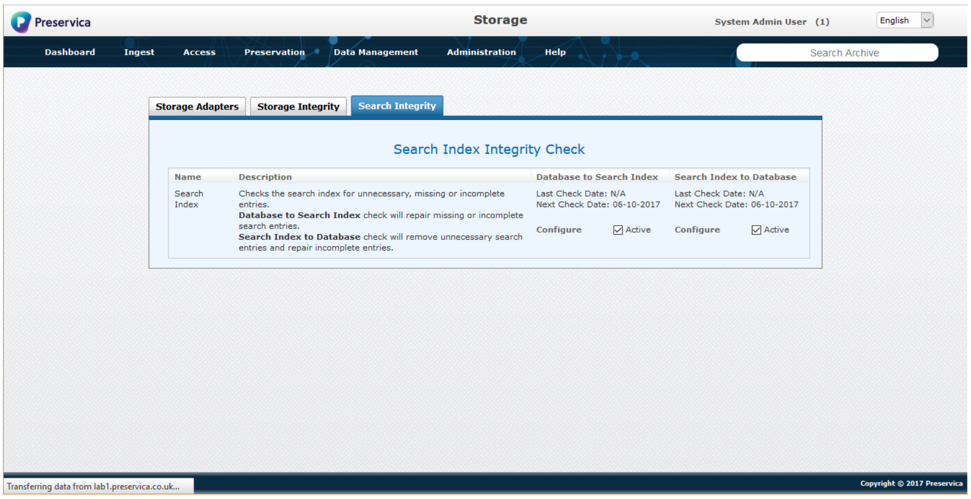
The Search Integrity tab displays a configuration table for the two checks to evaluate the integrity of search indexes associated with records within Preservica.

The first is the **Database to Search Index**. This iterates through records in the metadata store and compares the information there with the corresponding index entry, this will add any entries that are missing from the search index.

The second is the **Search Index to Database**. This iterates through records in the search index and compares the information there with the corresponding metadata store entry, this will remove any index entries that are not in the metadata store.

Where there are any differences in the entries from the metadata store and the index, both checks will repair the index using the information from the metadata store.

Figure 7.5. Search Integrity Tab



Search Index Integrity Check

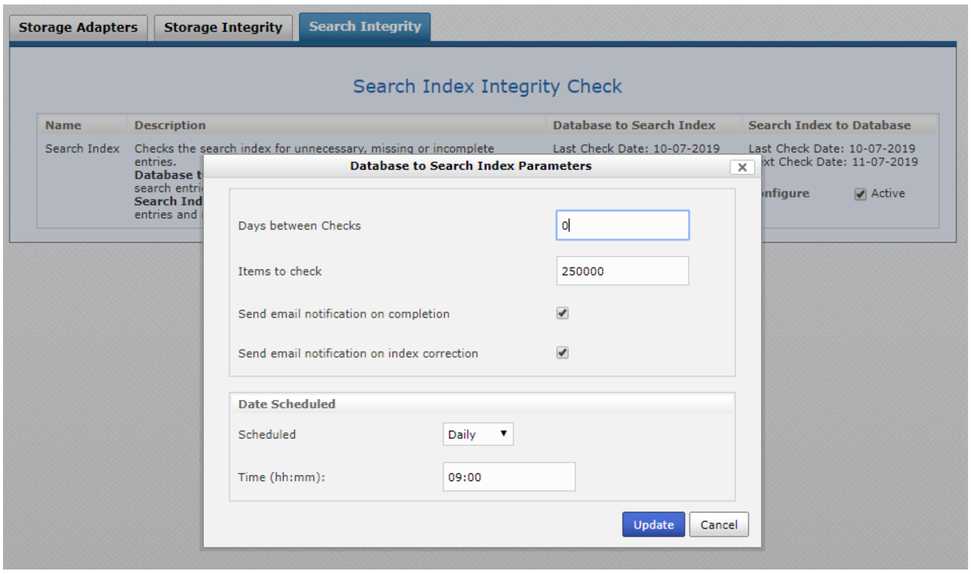
Name	Description	Database to Search Index	Search Index to Database
Search Index	Checks the search index for unnecessary, missing or incomplete entries. **Database to Search Index** check will repair missing or incomplete search entries. **Search Index to Database** check will remove unnecessary search entries and repair incomplete entries.	Last Check Date: N/A Next Check Date: 06-10-2017 Configure ☑ Active	Last Check Date: N/A Next Check Date: 06-10-2017 Configure ☑ Active

The **Database to Search Index** and **Search Index to Database** columns show information about the last date when each check was completed, and the dates of the next scheduled runs. They display whether a check is currently active (whether any checks will be triggered as scheduled), and provide a link to configure this check.

Upon selecting either link a pop-up dialogue will appear. The pop-up dialogues display, and allow the user to vary, a number of parameters which determine how each check runs, and how frequently:

- The **Days Between Checks** parameter will determine how frequently the same records will be checked.
- The **Items to check** parameter refers to how many items will be checked within every running of this check.
- Checking the **Send email notification on completion** option will result in an email being sent to the email set in the *integrity.manager.email* system property every time a search index check of this kind has been completed.
- Checking the **Send email notification on index correction** option will result in an email being sent to the user set in the *integrity.manager.email* system property when a check finds an inconsistency in the search index, and has corrected this. Note that if the **Send email notification on completion** is checked, this information will be received irrespective of whether this second option is checked.
- The **Date Scheduled** section allows the user to choose the rate at which the check is run (Daily, Weekly, Monthly, or Yearly) and select appropriate times, days of the week, days of the month, or months to correctly configure the selected option.

Figure 7.6. Amend Search Integrity Settings



Search Index Integrity Check

Name	Description	Database to Search Index	Search Index to Database
Search Index	Checks the search index for unnecessary, missing or incomplete entries. **Database t** search entri **Search Ind** entries and	Last Check Date: 10-07-2019	Last Check Date: 10-07-2019 xt Check Date: 11-07-2019 **nfigure** ☑ Active

Database to Search Index Parameters ✕

Days between Checks: 0

Items to check: 250000

Send email notification on completion ☑

Send email notification on index correction ☑

Date Scheduled

Scheduled: Daily ▾

Time (hh:mm): 09:00

Chapter 8. SIP Creator

See the SIP Creator User Guide [SCUG]

Chapter 9. Universal Access

See the Guide To Universal Access [GTUA]

Chapter 10. Custom Search Indexers

To provide efficient search facilities for ingested data, an Apache Solr-based, full-text search index is used by the Preservica archive (http://lucene.apache.org/solr/). By default, this indexes all the XIP fields on folders and assets, including a generic metadata field to allow text indexing of any descriptive metadata on entities. (Content objects and other sub-asset entities are not indexed separately, but information is made available on the asset's record.)

By default, Preservica will index most entity metadata, the entire descriptive metadata as tokenised text, and the full contents of certain file types (e.g. Office documents). This means that for standard search purposes, you won't need to add any custom indexing, even if you are using custom metadata schemas. For example, if you are using EAD metadata on your ingested material, people can still search (in Explorer and through Universal Access or other portals using API access) on terms in the EAD metadata, and the results will show those items through the default indexing of the metadata.

If you want your custom metadata fields to be usable as search filters within Explorer, or facets or filters in UA, then you will need to add a custom indexer for the schema you are using. Each indexer is a document that fulfils the built-in XML schema for custom search indexers, and defines four main items:

- The name of the schema (schemaName). This name is what appears in the Schema dropdown in Explorer.
- The URI of the schema (schemaUri). This should match the schemaUri on the metadata fragment, which will be the schema URI that you see on the XML Schemas tab of the Schema Management page for your custom metadata schema.
- A short name, which is used to refer to the schema in a more concise way (shortName). This name is usable as an XML namespace in the XPath expression for index terms.
- The term to add to the search index from this schema (terms). Each term (a <term> tag) has the following properties:
  - A name (indexName), which is used internally within the index. This should be a valid Solr identifier, i.e. no space, punctuation or special characters.
  - A display name (displayName), which is used in the Explorer filter dropdown and filter display.
  - The XPath expression to extract the indexed content from the metadata (xpath).
  - The type of the index (indexType). If omitted, this attribute defaults to STRING_DEFAULT; the possible values are:
    - STRING_DEFAULT, which specifies that the sdb_default query analyser will be used (by default this is Lucene's StandardAnalyzer, which tokenises the content);
    - STRING_EXACT, which will use the sdb_exact query analyser (which does not tokenise, so only an exact search match will produce a result);
    - STRING_FACET, which is similar to STRING_EXACT but preserves case, typically used for faceted search (queries against this field are case-sensitive in all searches);
    - DATE, which instructs Solr to index the field as a date and allows for date range querying (note that values in this field *must* satisfy ISO-8601 for date formats, and the time zone must be Z, e.g. `2016-06-15T02:43:46.000Z`);
    - LONG, which instructs Solr to index the field as a whole number, and allows for range querying.

- The XIP entity types (folder or asset) to which the indexer should be applied (xipTypes). This item is also optional; if it is not specified, the indexer will be applied to any item which has the relevant metadata defined. If it is, it should be a comma separated list of tokens from the list: SO, IO

- A boolean flag (true|false) to indicate if the field can be sorted (sortable): this attribute is optional and defaults to true (i.e. the field will be sortable)[1]. Note that if you do mark the field as sortable, Solr will mark the field as being *single*-valued; i.e. only one value will be stored in the index for this field for any Preservica entity. We recommend setting this value to true for any new indexers you create, unless the field needs to be multi-valued.

- A boolean flag (true|false) to indicate if the field can be used for faceting (facetable) : this attribute is optional and defaults to true (i.e. the field will be facetable).

A term with type STRING_FACET will not appear as a filter option in Explorer.

Changing the sortable flag on a term which already exists in the search index may cause problems, particularly for non-text field types.

If the index terms need to access more than one XML namespace, for example the OAI version of the Dublin Core metadata, then you can also specify one or more namespaceMapping elements with key and value attributes. These define namespace aliases that you can use in the XPath expressions for indices.

10.1. Example of Custom Indexing

For example, consider a custom metadata schema like this:

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns="http://fake.co.uk"
    xmlns:xs="http://www.w3.org/2001/XMLSchema"
    targetNamespace="http://www.preservica.com/
fakeExampleSchema" elementFormDefault="qualified">
    <xs:element name="Archivist">
        <xs:complexType>
            <xs:sequence>
                <xs:element name="Name" type="xs:string" />
                <xs:element name="Email" type="xs:string" />
            </xs:sequence>
        </xs:complexType>
    </xs:element>
</xs:schema>
```

In an XIP element, that will mean a metadata element that looks like:

```
<Metadata schemaURI="http://www.preservica.com/fakeExampleSchema">
    <Ref>5c2e587e-25c9-4000-a766-d3b9dccd8c00</Ref>
    <Entity>8a62c377-19d4-4602-b826-f28b80fd639a</Entity>
    <Content>
        <Archivist xmlns="http://www.preservica.com/fakeExampleSchema">
            <Name>A. Archivist</Name>
```

[1] In prior versions of Preservica, the default was non-sortable.

```
            <Email>a.archivist@myorg.com</Email>
        </Archivist>
    </Content>
</Metadata>
```

Now, imagine that we want to index the Name field as *Folder Owner* for folders, and *Asset Owner* for assets. We also want a sortable Owner field that we can use as a facet, or as a Content Metadata column in Universal Access. We'll index the Email field for everything, but only allow exact matches. Our indexer would look like:

```
<?xml version="1.0" encoding="UTF-8"?>
<index xmlns="http://www.preservica.com/customindex/v1">
    <schemaName>Example Custom Indexer</schemaName>
    <schemaUri>http://www.preservica.com/fakeExampleSchema</schemaUri>
    <shortName>fake</shortName>
    <term indexName="folder_owner" displayName="Folder Owner" xpath="//
fake:Name" xipTypes="SO" sortable="true"/>
    <term indexName="asset_owner" displayName="Asset Owner" xpath="//
fake:Name" xipTypes="IO" sortable="true"/>
    <term indexName="sortable_owner" displayName="Owner" xpath="//
fake:Name" indexType="STRING_FACET" sortable="true"/>
    <term indexName="email" displayName="Email" xpath="//
fake:Email" indexType="STRING_EXACT" sortable="true"/>
</index>
```

The Name field has been indexed as standard text so will be tokenised for search processes. The Email field has been indexed as an `STRING_EXACT` as you would be typically searching for a known email address.

In many cases you will be able to leave the xipTypes attribute off, as most metadata fields can be indexed on all types.

Indexers from Preservica v5 which have xipTypes set to v5 XIP object types will ignore the xipTypes attribute, i.e. those terms will be indexed for all entities in v6.

10.2. Indexing Dublin Core (OAI)

When a search indexer requires more than one XML namespace, you must specify namespaceMapping elements. The OAI wrapper of Dublin Core is a common case where this will be necessary. For example, to index the Title and Description fields from Dublin Core, you should index it as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<index xmlns="http://www.preservica.com/customindex/v1">
    <schemaName>DSpace OAI title/desc</schemaName>
    <schemaUri>http://www.openarchives.org/OAI/2.0/oai_dc/</schemaUri>
    <shortName>oai_dc</shortName>
    <term indexName="title" displayName="DC Title" xpath="//dc:title"/>
    <term indexName="desc" displayName="DC Description" xpath="//
dc:description"/>
    <namespaceMapping key="dc" value="http://purl.org/dc/elements/1.1/"/>
</index>
```

Note the OAI schema is the schemaUri of the indexer, and the dc namespace is included through a namespaceMapping tag.

You can use this mechanism for other custom metadata fragments which use multiple XML namespaces, for example if you have embedded XHTML or SVG.

10.3. Updating Indexers

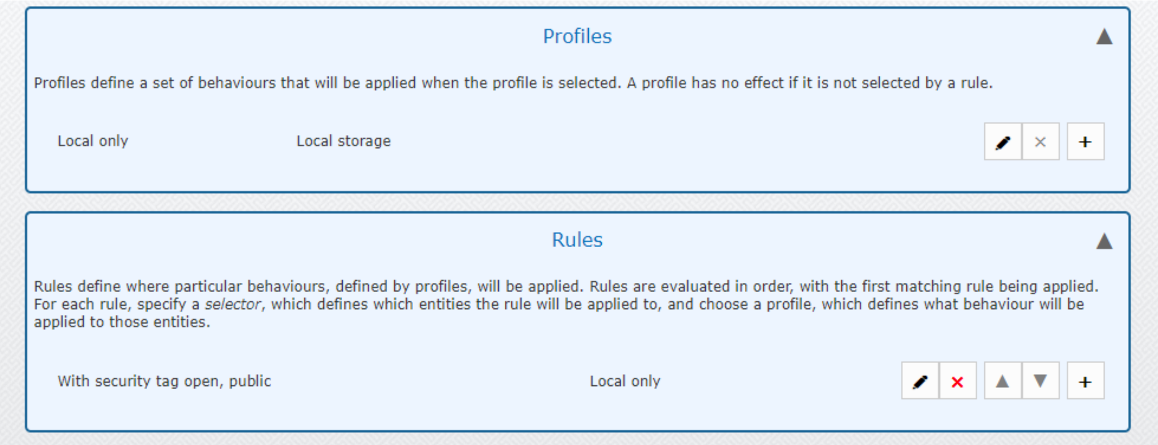
When you upload a custom indexer document (see Section 4.1.3 for details of how to do this), that indexer is immediately available to the system. That means that any new content ingested after that point which contains the relevant metadata will be indexed according to the indexer, and the schema and terms will be available as search filter options. If you want to update an existing indexer, remove and re-add it, and the changes will be made available.

However, no existing content will have the new or updated indexer, even if it has the relevant metadata. If you want the new indexer to apply to content which has already been ingested, you need to have that content re-indexed. The most convenient way to do that is to run the Re-Index workflow (a standard Data Management workflow).

Chapter 11. Settings and Policy Pages

Some configuration (for OCR, storage, and is done on settings pages. These pages all follow a common pattern of configuring profiles and rules.

Figure 11.1. The storage settings page



Profiles

Profiles define a set of behaviours that will be applied when the profile is selected. A profile has no effect if it is not selected by a rule.

Local only — Local storage

Rules

Rules define where particular behaviours, defined by profiles, will be applied. Rules are evaluated in order, with the first matching rule being applied. For each rule, specify a *selector*, which defines which entities the rule will be applied to, and choose a profile, which defines what behaviour will be applied to those entities.

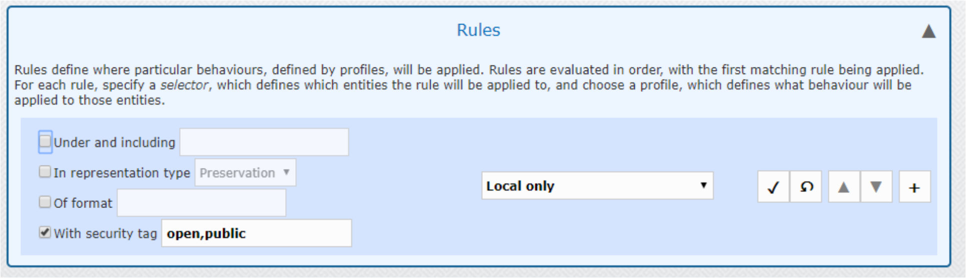
With security tag open, public — Local only

In the top section of the page, you configure **profiles**, which specify *what* to do. In this example it is the storage settings page and so the profile is configured to define storage routing. Each settings page has a different type of profile, appropriate to what is being configured.

In the lower part, you configure **rules**, which specify *where* a profile should be applied. Depending on the page, some or all of the following criteria can be set for a rule:

- A location in the hierarchy. Enter the reference of a folder. Content must be in that folder, or one of its descendants, to match.
- A representation type.
- A list of formats (PUIDs), as a comma separated list. Content must have a primary format in this list to match.
- A list of security tags, as a comma separated list. Content must have a tag in this list to match.

Figure 11.2. Editing a rule



Rules

Rules define where particular behaviours, defined by profiles, will be applied. Rules are evaluated in order, with the first matching rule being applied. For each rule, specify a *selector*, which defines which entities the rule will be applied to, and choose a profile, which defines what behaviour will be applied to those entities.

□ Under and including
□ In representation type: Preservation ▾
□ Of format
☑ With security tag: **open,public**

Local only ▾

If any of the criteria aren’t selected, they have no effect on whether content matches, and if none of them are set, all content will match the rule. In this example, content with the security tag *open* or *public* will match the rule, wherever it is and whatever format it is.

On the right side, select the profile which should be used for content that matches this rule.

11.1. How Rules are Evaluated

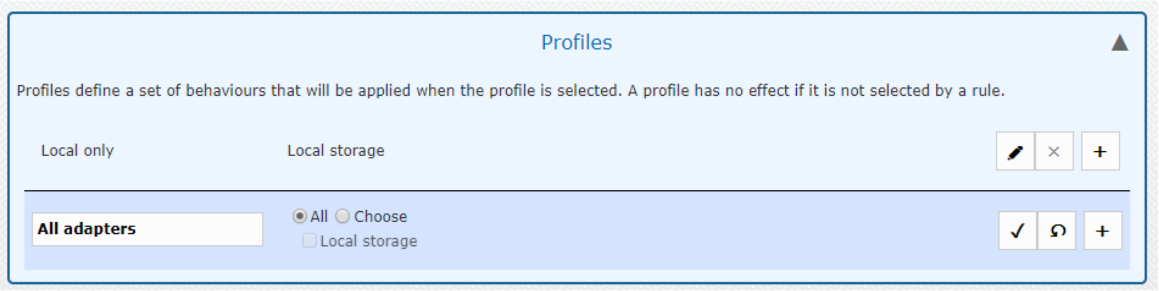
When performing an action that is configured in this way, Preservica will first select the generation of content that it will do the action on. (For storage, it will evaluate every generation, as everything should be stored; for other purposes the latest active generation of a representation is used.)

Then, each rule in turn will be evaluated according to the criteria selected for that rule. If it matches, the profile set for that rule will be used, and the actions it defines will be done (in the storage example above, the content will only be stored on the local adapter). Rules are evalated in order, and the first rule that matches will be applied, so you should order your rules with the most specific ones first. Use the arrow buttons to re-order rows.

If no rules match the selected content, no profile will be applied, and the default action will take place.

Rule evaluation is done for every piece of content independently, so content within the same ingest package or indexing operation could select different rules and apply different profiles.

11.2. Storage Settings



Profiles

Profiles define a set of behaviours that will be applied when the profile is selected. A profile has no effect if it is not selected by a rule.

Local only   Local storage

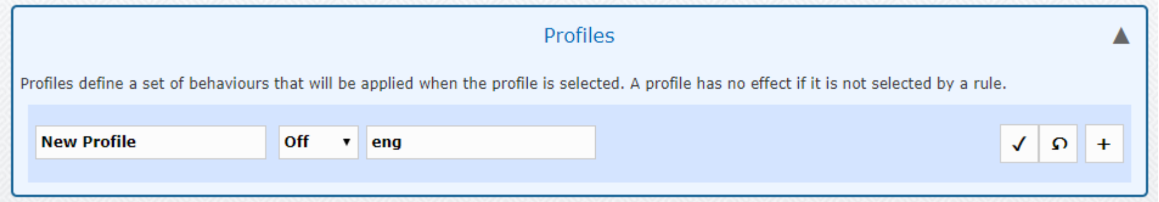
All adapters   ● All ○ Choose
□ Local storage

Storage profiles define which adapters content will be stored to. You can specify that content should be stored to all adapters, or select particular adapters.

Note that adapters still appear here if they are currently read-only or unavailable. However, if you attempt to run an ingest workflow and the profile selected for some content doesn't route that content to any writeable adapters, the ingest will fail.

The default action is to store content on all adapters.

11.3. OCR Settings



Profiles

Profiles define a set of behaviours that will be applied when the profile is selected. A profile has no effect if it is not selected by a rule.

New Profile   Off ▾   eng

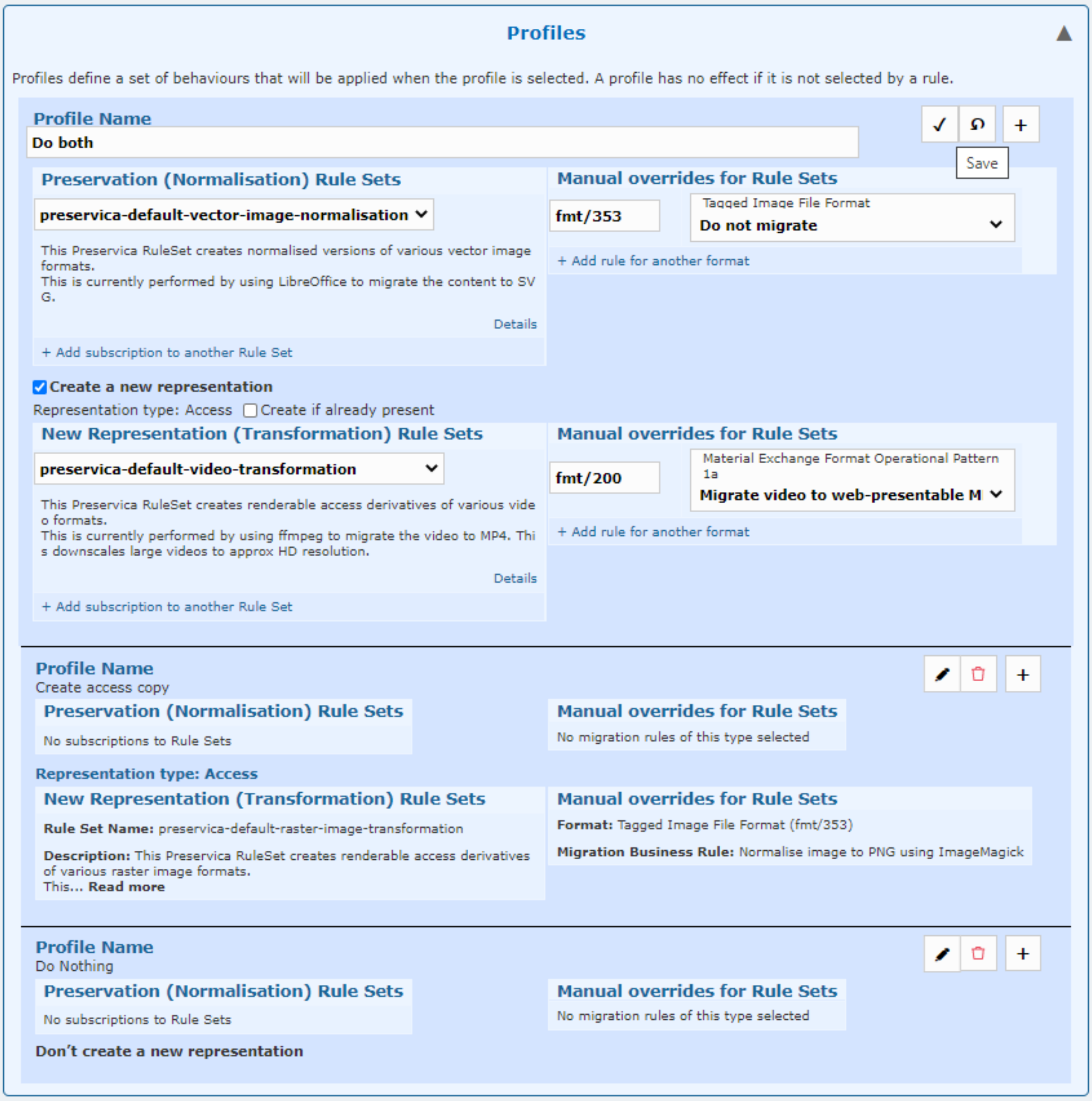
OCR profiles define whether and how OCR is run against content when indexing. The first option is where the output from OCR should go – into the search index, or off. Optionally you may specify languages, as a comma separated list of 3 letter langauge codes (e.g. *eng,deu*). These will be used as a hint to Tesseract.

These profiles are evaluated when performing full text indexing, either after ingesting new content or when running a re-index workflow with the full text option set.

The rule selector won't let you specify a representation type, because the representation used for full text search indexing is the only one eligible for OCR.

The default action is to not do OCR on anything.

11.4. Migration Settings



Profiles

Profiles define a set of behaviours that will be applied when the profile is selected. A profile has no effect if it is not selected by a rule.

Profile Name
Do both

Preservation (Normalisation) Rule Sets

preservica-default-vector-image-normalisation

This Preservica RuleSet creates normalised versions of various vector image formats.
This is currently performed by using LibreOffice to migrate the content to SV G.

Details

+ Add subscription to another Rule Set

Manual overrides for Rule Sets

fmt/353 — Tagged Image File Format: Do not migrate

+ Add rule for another format

☑ Create a new representation

Representation type: Access ☐ Create if already present

New Representation (Transformation) Rule Sets

preservica-default-video-transformation

This Preservica RuleSet creates renderable access derivatives of various vide o formats.
This is currently performed by using ffmpeg to migrate the video to MP4. Thi s downscales large videos to approx HD resolution.

Details

+ Add subscription to another Rule Set

Manual overrides for Rule Sets

fmt/200 — Material Exchange Format Operational Pattern 1a: Migrate video to web-presentable M

+ Add rule for another format

Profile Name
Create access copy

Preservation (Normalisation) Rule Sets

No subscriptions to Rule Sets

Manual overrides for Rule Sets

No migration rules of this type selected

Representation type: Access

New Representation (Transformation) Rule Sets

Rule Set Name: preservica-default-raster-image-transformation

Description: This Preservica RuleSet creates renderable access derivatives of various raster image formats.
This... **Read more**

Manual overrides for Rule Sets

Format: Tagged Image File Format (fmt/353)

Migration Business Rule: Normalise image to PNG using ImageMagick

Profile Name
Do Nothing

Preservation (Normalisation) Rule Sets

No subscriptions to Rule Sets

Manual overrides for Rule Sets

No migration rules of this type selected

Don't create a new representation

Migration profiles control automatic migration of content. This is applied after ingest automatically, and can be applied by individual *Preserve* or *Create New Representation* workflows by configuring the workflow context to tick the “Use Migration Settings” option. In each profile you can specify normalisation (i.e. for preservation purposes) business rules, or request the creation of a new representation with migration business rules, or both. Use the checkbox to turn the representation creation section on or off.

For each type of migration, you can select to subscribe to pre-defined Rule Sets. Each Rule Set provides migration rules for a range of formats. Each Rule Set typically provides formats with a common theme, for example, all document formats, or all image formats, or all formats that a particular institution suggests should be migrated. Each Rule Set that you select represents a “subscription” in the sense that the Rule Set is evaluated each time it is used (i.e. during ingest/migration workflows), and any changes that have been made will be applied automatically. You are choosing the follow the high level intention of the Rule Set rather than the exact implementation at any given time. Rule Sets can be selected from the drop-down when you select *Add subscription to another Rule Set*.

You can additionally set manual overrides to the Rule Sets for specific formats in the tables to the right hand side of the page. The rules that are available for a particular format come from the Registry. In the *normalisation* table, only rules applicable to normalisation will be shown. When you select *Add rule from another format*, enter the format PUID in the text field, which will then populate a dropdown to allow you to select the business rule.

You can’t select formats in the main rule selector on this page because format criteria are replaced by the migration rules in the profiles.

The default action is to do no migrations.

Preservica
Active digital preservation

www.preservica.com

Reference: git/doc/UG/SAG | Issue: 6.6.1

Preservica, 32 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire, OX14 3YS, UK info@preservica.com - preservica.com