Maintaining Matterhorn Disk Space

Users running Avalon Media System for some length of time might notice an increase in the disk space that Matterhorn consumes. Matterhorn is the transcoding workflow engine used by Avalon, and the loss of disk space is a result of it holding onto temporary workflow artifacts indefinitely—even when they are no longer needed. This is especially true for Avalon versions through 3.2. In release 3.3, changes were made to reduce this problem.

Although the need has been reduced due to workflow changes, there still may be times when it is desirable to clean up Matterhorn's work directories. The script below can be run to accomplish this task. It performs three main tasks:

1. It removes temporary artifacts related to any workflow that was successfully completed by the previous day.
2. It removes temporary artifacts related to any workflow that Matterhorn no longer knows about.
3. It removes temporary artifacts that were saved for failed workflows.

This script should be edited as necessary for individual installations.

Deleting the workflow may leave job in a bad state

```
2016-10-19 12:59:00 WARN (WorkflowServiceImpl:1683) - Exception while accepting job Job {id:3502, version:
25}
org.opencastproject.util.NotFoundException: Workflow '3494' has been deleted
at org.opencastproject.workflow.impl.WorkflowServiceImpl.getWorkflowById(WorkflowServiceImpl.java:480)
at org.opencastproject.workflow.impl.WorkflowServiceImpl.process(WorkflowServiceImpl.java:1659)
at org.opencastproject.workflow.impl.WorkflowServiceImpl$JobRunner.call(WorkflowServiceImpl.java:2048)
at org.opencastproject.workflow.impl.WorkflowServiceImpl$JobRunner.call(WorkflowServiceImpl.java:2014)
at java.util.concurrent.FutureTask$Sync.innerRun(FutureTask.java:303)
at java.util.concurrent.FutureTask.run(FutureTask.java:138)
at java.util.concurrent.ThreadPoolExecutor$Worker.runTask(ThreadPoolExecutor.java:886)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:908)
at java.lang.Thread.run(Thread.java:662)
2016-10-19 12:59:00 WARN (WorkflowServiceImpl:1689) - Unable to parse workflow instance
org.opencastproject.util.NotFoundException: Workflow '3494' has been deleted
at org.opencastproject.workflow.impl.WorkflowServiceImpl.getWorkflowById(WorkflowServiceImpl.java:480)
at org.opencastproject.workflow.impl.WorkflowServiceImpl.process(WorkflowServiceImpl.java:1659)
at org.opencastproject.workflow.impl.WorkflowServiceImpl$JobRunner.call(WorkflowServiceImpl.java:2048)
at org.opencastproject.workflow.impl.WorkflowServiceImpl$JobRunner.call(WorkflowServiceImpl.java:2014)
at java.util.concurrent.FutureTask$Sync.innerRun(FutureTask.java:303)
at java.util.concurrent.FutureTask.run(FutureTask.java:138)
at java.util.concurrent.ThreadPoolExecutor$Worker.runTask(ThreadPoolExecutor.java:886)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:908)
at java.lang.Thread.run(Thread.java:662)
```

```
#!/bin/bash

USERNAME='matterhorn_system_account'
PASSWD='CHANGE_ME'
MATTERHORN_HOME='/usr/local/matterhorn/work/content'
YESTERDAY=`date --date="yesterday" -u +"%Y-%m-%dT00:00:00Z"`

cleanupDir() {
  mpdir=$1
  for mp in `ls ${mpdir}`; do
    #Delete Mediapackages that are done processing
    curl --digest -u "${USERNAME}:${PASSWD}" -H "X-Requested-Auth: Digest" -H "X-Opencast-Matterhorn-
Authorization: true" "http://localhost:18080/workflow/instances.json?state=SUCCEEDED%2C%20STOPPED%2C%20SKIPPED%
2C%20FAILED&mp=${mp}&todate=${YESTERDAY}" 2>/dev/null | grep -qci -e '"totalCount":"1"' && echo "Deleting
${mpdir}/${mp}" && rm -r ${mpdir}/${mp}
    #Delete Mediapackages that Matterhorn doesn't know about anymore
    curl --digest -u "${USERNAME}:${PASSWD}" -H "X-Requested-Auth: Digest" -H "X-Opencast-Matterhorn-
Authorization: true" "http://localhost:18080/workflow/instances.json?mp=${mp}" 2>/dev/null | grep -qci -e
'"totalCount":"0"' && echo "Deleting ${mpdir}/${mp}" && rm -r ${mpdir}/${mp}
  done
}

#Delete Mediapackages in three possible locations
cleanupDir "${MATTERHORN_HOME}/files/mediapackage"
cleanupDir "${MATTERHORN_HOME}/workspace/mediapackage"
cleanupDir "${MATTERHORN_HOME}/archive-temp"

#Remove zips from failed workflows
rm ${MATTERHORN_HOME}/files/collection/failed.zips/*
rm ${MATTERHORN_HOME}/workspace/collection/failed.zips/*
```

Cleaning Up the Matterhorn Database

Matterhorn will query its own `mh_job` table every minute or so, and since older jobs are not removed from this table it will eventually reach a point where these queries become a noticeable drain on system resources. To clean out the table:

1. Locate your matterhorn install and check etc/config.properties for the database you are using (default is /usr/local/matterhorn/etc/config.properties)
2. Open up that database via your browser of choice and select the `mh_job` table
3. Delete rows where the operation value is **NOT** START_WORKFLOW (we want to retain those jobs for later reference). You can use `date_created` to scope this delete via time.
4. Your db table should be much smaller now.

Example

```
delete from mh_job where operation <> 'START_WORKFLOW' and date_created < '2018-12-01';
```