

Scheduled Tasks via Cron

Several DSpace features **require** that a script is run regularly (via cron, or similar). Some of these features include:

- the [e-mail subscription feature](#) that alerts users of new items being deposited;
- the ['media filter' tool](#), that generates thumbnails of images and extracts the full-text of documents for indexing;
- the ['checksum checker'](#) that tests the bitstreams in your repository for corruption;
- the [sitemap generator](#), which enhances the ability of major search engines to index your content and make it findable;
- the [curation system queueing feature](#), which allows administrators to "queue" tasks (to run at a later time) from the Admin UI;
- and [Discovery](#) (search & browse), [OAI-PMH](#) and [Usage Statistics](#) all receive performance benefits from regular re-optimization.

These regularly scheduled tasks should be setup via either [cron](#) (for Linux/Mac OSX) or [Windows Task Scheduler](#) (for Windows).

Recommended Cron Settings

If you are on Linux or Mac OSX, you should add these cron settings under the OS account which is running DSpace (or Tomcat). For example, login as that user and type the following to edit the user's crontab.

```
crontab -e
```

While every DSpace installation is unique, in order to get the most out of DSpace, we highly recommend enabling these basic cron settings (the settings are described in the comments):

```
## SAMPLE CRONTAB FOR A PRODUCTION DSPACE
## You obviously may wish to tweak this for your own installation,
## but this should give you an idea of what you likely wish to schedule via cron.
##
## NOTE: You may also need to add additional sysadmin related tasks to your crontab
## (e.g. zipping up old log files, or even removing old logs, etc).

#-----
# GLOBAL VARIABLES
#-----
# Full path of your local DSpace Installation (e.g. /home/dspace or /dspace or similar)
# MAKE SURE TO CHANGE THIS VALUE!!!
DSPACE = [dspace]

# Shell to use
SHELL=/bin/sh

# Add all major 'bin' directories to path
PATH=/usr/local/sbin:/usr/local/bin:/sbin:/bin:/usr/sbin:/usr/bin

# Set JAVA_OPTS with defaults for DSpace Cron Jobs.
# Only provides 512MB of memory by default (which should be enough for most sites).
JAVA_OPTS="-Xmx512M -Xms512M -Dfile.encoding=UTF-8"

#-----
# HOURLY TASKS (Recommended to be run multiple times per day, if possible)
# At a minimum these tasks should be run daily.
#-----

# Regenerate DSpace Sitemaps every 8 hours (12AM, 8AM, 4PM).
# SiteMaps ensure that your content is more findable in Google, Google Scholar, and other major search engines.
0 0,8,16 * * * $DSPACE/bin/dspace generate-sitemaps > /dev/null

#-----
# DAILY TASKS
# (Recommended to be run once per day. Feel free to tweak the scheduled times below.)
#-----

# Update the OAI-PMH index with the newest content (and re-optimize that index) at midnight every day
# NOTE: ONLY NECESSARY IF YOU ARE RUNNING OAI-PMH
# (This ensures new content is available via OAI-PMH and ensures the OAI-PMH index is optimized for better
performance)
0 0 * * * $DSPACE/bin/dspace oai import -o > /dev/null
```

```

# Note, you may want to consider rebuilding your entire OAI index every night... it doesn't take very long, and
it ensures the feed is current
# to do so, just comment out the oai import line above and uncomment the one below
# 0 23 * * * dspace $DSpace/bin/dspace oai import -c -o clean-cache

# Clean and Update the Discovery indexes at midnight every day
# (This ensures that any deleted documents are cleaned from the Discovery search/browse index)
0 0 * * * $DSpace/bin/dspace index-discovery > /dev/null

# Re-Optimize the Discovery indexes at 12:30 every day
# (This ensures that the Discovery Solr Index is re-optimized for better performance)
30 0 * * * $DSpace/bin/dspace index-discovery -o > /dev/null

# Cleanup Web Spiders from DSpace Statistics Solr Index at 01:00 every day
# NOTE: ONLY NECESSARY IF YOU ARE RUNNING SOLR STATISTICS
# (This removes any known web spiders from your usage statistics)
0 1 * * * $DSpace/bin/dspace stats-util -i

# Re-Optimize DSpace Statistics Solr Index at 01:30 every day
# NOTE: ONLY NECESSARY IF YOU ARE RUNNING SOLR STATISTICS
# (This ensures that the Statistics Solr Index is re-optimized for better performance)
30 1 * * * $DSpace/bin/dspace stats-util -o

# Send out subscription e-mails at 02:00 every day
# (This sends an email to any users who have "subscribed" to a Collection, notifying them of newly added
content.)
0 2 * * * $DSpace/bin/dspace sub-daily

# Run the media filter at 03:00 every day.
# (This task ensures that thumbnails are generated for newly add images,
# and also ensures full text search is available for newly added PDF/Word/PPT/HTML documents)
0 3 * * * $DSpace/bin/dspace filter-media

# Run any Curation Tasks queued from the Admin UI at 04:00 every day
# (Ensures that any curation task that an administrator "queued" from the Admin UI is executed
# asynchronously behind the scenes)
0 4 * * * $DSpace/bin/dspace curate -q admin_ui

#-----
# WEEKLY TASKS
# (Recommended to be run once per week, but can be run more or less frequently, based on your local needs
/policies)
#-----
# Run the checksum checker at 04:00 every Sunday
# By default it runs through every file (-l) and also prunes old results (-p)
# (This re-verifies the checksums of all files stored in DSpace. If any files have been changed/corrupted,
checksums will differ.)
0 4 * * * $DSpace/bin/dspace checker -l -p
# NOTE: LARGER SITES MAY WISH TO USE DIFFERENT OPTIONS. The above "-l" option tells DSpace to check
*everything*.
# If your site is very large, you may need to only check a portion of your content per week. The below
commented-out task
# would instead check all the content it can within *one hour*. The next week it would start again where it
left off.
#0 4 * * 0 $DSpace/bin/dspace checker -d 1h -p

# Mail the results of the checksum checker (see above) to the configured "mail.admin" at 05:00 every Sunday.
# (This ensures the system administrator is notified whether any checksums were found to be different.)
0 5 * * 0 $DSpace/bin/dspace checker-emailer

#-----
# MONTHLY TASKS
# (Recommended to be run once per month, but can be run more or less frequently, based on your local needs
/policies)
#-----
# Permanently delete any bitstreams flagged as "deleted" in DSpace, on the first of every month at 01:00
# (This ensures that any files which were deleted from DSpace are actually removed from your local filesystem.
# By default they are just marked as deleted, but are not removed from the filesystem.)
0 1 1 * * $DSpace/bin/dspace cleanup > /dev/null

#-----

```

```
# YEARLY TASKS (Recommended to be run once per year)
#-----
# At 2:00AM every January 1, "shard" the DSpace Statistics Solr index.
# This ensures each year has its own Solr index, which improves performance.
# NOTE: ONLY NECESSARY IF YOU ARE RUNNING SOLR STATISTICS
# NOTE: This is scheduled here for 2:00AM so that it happens *after* the daily cleaning & re-optimization of
this index.
0 2 1 1 * $DSPACE/bin/dspace stats-util -s
```