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 (optional) ‘checksum checker’ that tests the bitstreams in your repository for corruption;
- and the (optional) registration of DOIs using DataCite as registration agency.

There are some optional periodic tasks as well:

- Updating the geolocation database used to enrich usage statistics. At this writing, the database publisher issues monthly updates.

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 Tomcat (and owns the [dspace] installation directory). 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):

```bash
## 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)
#--------------
# Send information about new and changed DOIs to the DOI registration agency
# NOTE: ONLY NECESSARY IF YOU REGISTER DOIS USING DATAcite AS REGISTRATION AGENCY (Disabled by default)
# 0 4,12,20 * * * $DSPACE/bin/dspace doi-organiser -u -q ; $DSPACE/bin/dspace doi-organiser -s -q ; $DSPACE/bin/dspace doi-organiser -r -q ; $DSPACE/bin/dspace doi-organiser -d -q
#----------------
# 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 at midnight every day
# REQUIRED to update content available in OAI-PMH (However, it can be removed if you do not enable OAI-PMH)
0 0 * * * $DSPACE/bin/dspace oai import > /dev/null
```
# Clean and Update the Discovery indexes at midnight every day
# (This ensures that any deleted documents are cleaned from the Discovery search/browse index)
# RECOMMENDED to ensure your search/browse index stays fresh.
0 0 * * * $DSPACE/bin/dspace index-discovery > /dev/null

# run the index-authority script once a day at 12:45 to ensure the Solr Authority cache is up to date
45 0 * * * $DSPACE/bin/dspace index-authority > /dev/null

# Cleanup Web Spiders from DSpace Statistics Solr Index at 01:00 every day
# (This removes any known web spiders from your usage statistics)
# RECOMMENDED if you are running Solr Statistics.
0 1 * * * $DSPACE/bin/dspace stats-util -i

# Send out "daily" update subscription e-mails at 02:00 every day
# (This sends an email to any users who have "subscribed" to a Community/Collection, notifying them of newly added content.)
# REQUIRED for daily "Email Subscriptions" to work properly.
0 2 * * * $DSPACE/bin/dspace subscription-send -f D

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

#----------------
# WEEKLY TASKS
# (Recommended to be run once per week, but can be run more or less frequently, based on your local needs/policies)
#----------------
# Send out "weekly" update subscription e-mails at 02:00 every Sunday
# (This sends an email to any users who have "subscribed" to a Community/Collection, notifying them of newly added content.)
# REQUIRED for weekly "Email Subscriptions" to work properly.
0 2 * * 0 $DSPACE/bin/dspace subscription-send -f W

# 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.)
# OPTIONAL, but useful if you want to enable regular regular checksum validation of files stored in DSpace.
# 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)
#----------------
# Send out "monthly" update subscription e-mails at 02:00, on the first of every month
# (This sends an email to any users who have "subscribed" to a Community/Collection, notifying them of newly added content.)
# REQUIRED for monthly "Email Subscriptions" to work properly.
0 2 1 * * $DSPACE/bin/dspace subscription-send -f M

# 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.)
# REQUIRED to fully remove deleted content files from the "assetstore" folder
0 1 1 * * $DSPACE/bin/dspace cleanup > /dev/null