

DSpace 8 - Improvements to Processes

- [The create process page](#)
- [The overview page](#)
- [Bulk deletion](#)

The create process page

New Process

https://demo.dspace.org/processes/new

Name

filter-media

Script

filter-media

export
Batch Export to Simple Archive Format (SAF)

filter-media
Perform the media filtering to extract full text from documents and to create thumbnails

harvest
Manage the OAI-PMH harvesting of external collections

import
Batch Import from Simple Archive Format (SAF)

index-discovery
Update Discovery Solr Search Index

metadata-deletion
Delete all the values of the specified metadata field

filter-media
Perform the media filtering to extract full text from documents and to create thumbnails

-f --force force all bitstreams to be processed

-h --help help

-i --id <string> ONLY process bitstreams belonging to identifier

-m --max <number> process no more than maximum items

-p --plugins <string> ONLY run the specified Media Filter plugin(s) listed from 'filter.plugins' in dspace.cfg. Separate multiple with a comma (,) (e.g. MediaFilterManager -p "Word Text Extractor", "PDF Text Extractor")

-q --quiet do not print anything except in the event of errors.

-s --skip <string> SKIP the bitstreams belonging to identifier Separate multiple identifiers with a comma (,) (e.g. MediaFilterManager -s 123456789/34,123456789/323)

-v --verbose print all extracted text and other details to STDOUT

We've heard from clients that they often don't know what a script does simply by the script name. This is a problem when starting a new process, as well as on the overview page

So we propose two changes here.

First, show the description for a script underneath the script name in the script select. These descriptions already exist, but may need to be improved for some scripts

Second, add a name field for a process.

New Process
https://demo.dspace.org/processes/new

New Process

Name

filter-media -f -i 123456789/1234

Script

filter-media

Parameters

--force

--id

123456789/1234

Add a parameter

filter-media

Perform the media filtering to extract full text from documents and to create thumbnails

-f --force

force all bitstreams to be processed

-h --help

help

-i --id <string>

ONLY process bitstreams belonging to identifier

-m --max <number>

process no more than maximum items

-p --plugins <string>

ONLY run the specified Media Filter plugin(s) listed from 'filter.plugins' in dspace.cfg. Separate multiple with a comma (,) (e.g. MediaFilterManager -p "Word Text Extractor","PDF Text Extractor")

-q --quiet

do not print anything except in the event of errors.

-s --skip <string>

SKIP the bitstreams belonging to identifier Separate multiple identifiers with a comma (,) (e.g. MediaFilterManager -s 123456789/34,123456789/323)

-v --verbose

print all extracted text and other details to STDOUT

Initially the UI will automatically update this Name input, with the script name, followed by the parameters

New Process
https://demo.dspace.org/processes/new

New Process

Name

Regenerate thumbnails for publications

Script

filter-media

Parameters

--force

--id

123456789/1234

Add a parameter

filter-media

Perform the media filtering to extract full text from documents and to create thumbnails

-f --force

force all bitstreams to be processed

-h --help

help

-i --id <string>

ONLY process bitstreams belonging to identifier

-m --max <number>

process no more than maximum items

-p --plugins <string>

ONLY run the specified Media Filter plugin(s) listed from 'filter.plugins' in dspace.cfg. Separate multiple with a comma (,) (e.g. MediaFilterManager -p "Word Text Extractor","PDF Text Extractor")

-q --quiet

do not print anything except in the event of errors.

-s --skip <string>

SKIP the bitstreams belonging to identifier Separate multiple identifiers with a comma (,) (e.g. MediaFilterManager -s 123456789/34,123456789/323)

-v --verbose

print all extracted text and other details to STDOUT

But the user can overwrite that name, and once they do it will no longer change automatically.

The overview page

The screenshot shows a web application titled "Processes Overview" with a search bar and a "Go" button. Below the search bar is a "Start a new process" button. The left sidebar contains filters for Status (Running, Queued, Succeeded, Failed), Started by (Anybody, Me, User: Donald Smith Jr.), and Script (filter-media). A "Reset all filters" button is at the bottom of the filters. The main content area displays three tables: Running, Queue, and Completed. Each table has columns for Name, Started by, and Time elapsed/Completed on. The Running table shows one process: filter-media -f. The Queue table shows two processes: index-discovery -b and metadata-import -f import.csv. The Completed table shows five processes: curate -t profileformats -i all, harvest -g -i com 20.500.12542 184, metadata-export -i 746fe783-2509-4b35-901d-094d, filter-media -f, and filter-media -f. Each process row includes a status icon and a trash icon for deletion. A pagination bar at the bottom shows page 1 of 10.

Status	Name	Started by	Time elapsed	Actions
Running	filter-media -f	Donald Smith Jr.	2m 31s	[Status Icon] [Trash Icon]

Status	Name	Started by	Added on	Actions
Queue	index-discovery -b	James Howard	2023-10-25 11:33:1	[Status Icon] [Trash Icon]
Queue	metadata-import -f import.csv	Elissa Stevenson	2023-10-25 12:02:4	[Status Icon] [Trash Icon]

Status	Name	Started by	Completed on	Actions
Completed	curate -t profileformats -i all	Doyle Waller	2023-10-24 16:49:0	[Status Icon] [Trash Icon]
Completed	harvest -g -i com 20.500.12542 184	Donald Smith Jr.	2023-10-24 16:03:0	[Status Icon] [Trash Icon]
Completed	metadata-export -i 746fe783-2509-4b35-901d-094d	Joseph Banks	2023-10-23 13:49:4	[Status Icon] [Trash Icon]
Completed	filter-media -f	Donald Smith Jr.	2023-10-19 14:12:0	[Status Icon] [Trash Icon]
Completed	filter-media -f	Donald Smith Jr.	2023-10-19 14:05:3	[Status Icon] [Trash Icon]

The goal here was to make it possible to search and filter processes, similar to the way we do for items in discovery. This makes it possible to only show processes with a certain status, or for a certain script or started by a certain user

The page is divided into 3 lists, because it gives you a better overview at a glance of the status, and because the relevant information for a process changes depending on its status. E.g for a running process it's important to know how long it's been running. For a queued process when it was added, and for a completed one, when it completed.

Each process also gets an icon to indicate its status at a glance

Each of these lists will have separate pagination, however the default page size will be large enough (e.g. 10) that in practice you usually won't see more than one page for Running and Queued

We add the ability to remove running and queued processes. We also investigated the ability to reorder queued processes, but this looks like it's quite difficult to do on the backend, so we'll keep that out for now

Running processes will be polled using the method contributed in #2480, when a process completes all 3 lists will be re-retrieved automatically

The name column shows that process name property mentioned above

Processes Overview

← → ↻

https://demo.dspace.org/processes?f.user=Donald%20Smith%20Jr.&query=filter

≡

➤

⊕

✎

↶

↷

🔑

🔍

☰

🔿

📄

⚙️

➤➤

Processes Overview

filter

Go

Start a new process

Status

☒ Running

☒ Queued

☒ Succeeded

☒ Failed

Started by

Anybody

Me

User:

Donald Smith Jr.

Script

☐ filter-media

Reset all filters

Running

	Name	Started by	Time elapsed	
⚙️	filter-media -f	Donald Smith Jr.	2m 31s	🗑️

Queue

No matching processes

Completed

	Name	Started by	Completed on	
✅	filter-media -f	Donald Smith Jr.	2023-10-19 14:12:0	🗑️
❗	filter-media -f	Donald Smith Jr.	2023-10-19 14:05:3	🗑️

Here you see the page filtered using both a user name, and the search query "filter"

Bulk deletion

We also propose to include a script that will delete completed processes older than a given period. We discussed creating a custom endpoint for bulk deletion, but using a script has the advantage that it can be scheduled in a crontab (e.g. you schedule to run the script every day, to delete all processes older than 30 days), while it can still be run as a process from the UI too