Mediafilters for Transforming DSpace Content

1 MediaFilters: Transforming DSpace Content
- 1.1 Overview
- 1.2 Available Media Filters
- 1.3 Enabling/Disabling MediaFilters
- 1.4 Executing (via Command Line)
- 1.5 Creating Custom MediaFilters
  - 1.5.1 Creating a simple Media Filter
  - 1.5.2 Creating a Dynamic or "Self-Named" Format Filter
- 1.6 Configuration parameters

MediaFilters: Transforming DSpace Content

Overview

DSpace can apply filters or transformations to files/bitstreams, creating new content. Filters are included that extract text for **full-text searching**, and create **thumbnails** for items that contain images. The media filters are controlled by the `dspace filter-media` script which traverses the asset store, invoking all configured `MediaFilter` or `FormatFilter` classes on files/bitstreams (see Configuring Media Filters for more information on how they are configured).

Available Media Filters

Below is a listing of all currently available Media Filters, and what they actually do:

Name	Java Class	Function	Default input formats	Enabled by Default?
Text Extractor (7.3 or above)	`org.dspace.app.mediafilter.TikaTextExtractionFilter`	As of 7.3, all text extraction for Full text indexing takes place in a single filter. This filter uses Apache Tika which supports a wide variety of formats (e.g. Microsoft products, PDF, HTML, Text, etc). Additional formats may be configured from the Tika supported formats list at https://tika.apache.org/2.3.0/formats.html	Adobe PDF, Microsoft formats (Word, PPT, Excel), CSV, HTML, RTF, Text, OpenDocument formats (Text, Spreadsheet, Presentation)	yes
PDF Text Extractor (7.2 or below)	`org.dspace.app.mediafilter.PDFFilter`	extracts the full text of Adobe PDF documents (only if text-based or OCRed) for full text indexing. (Uses the Apache PDFBox tool)	Adobe PDF	yes
HTML Text Extractor (7.2 or below)	`org.dspace.app.mediafilter.HTMLFilter`	extracts the full text of HTML documents for full text indexing. (Uses Swing's HTML Parser)	HTML, Text	yes
Word Text Extractor (7.2 or below)	`org.dspace.app.mediafilter.PoiWordFilter`	extracts the full text of Microsoft Word and Microsoft Word XML documents for full text indexing. (Uses the "Apache POI" tools.)	Microsoft Word, Microsoft Word XML	yes
Excel Text Extractor (7.2 or below)	`org.dspace.app.mediafilter.ExcelFilter`	extracts the full text of Microsoft Excel documents for full text indexing. (Uses the "Apache POI" tools.)	Microsoft Excel, Microsoft Excel XML	yes
PowerPoint Text Extractor (7.2 or below)	`org.dspace.app.mediafilter.PowerPointFilter`	extracts the full text of slides and notes in Microsoft PowerPoint and PowerPoint XML documents for full text indexing. (Uses the Apache POI tools.)	Microsoft Powerpoint, Microsoft Powerpoint XML	yes
PDFBox JPEG Thumbnail	`org.dspace.app.mediafilter.PDFBoxThumbnail`	creates thumbnail images of the first page of PDF files	Adobe PDF	yes
JPEG Thumbnail	`org.dspace.app.mediafilter.JPEGFilter`	creates thumbnail images of GIF, JPEG and PNG files	BMP, GIF, JPEG, image/png	yes
Branded Preview JPEG	`org.dspace.app.mediafilter.BrandedPreviewJPEGFilter`	creates a branded preview image for GIF, JPEG and PNG files	BMP, GIF, JPEG, image/png	no

ImageMagick Image Thumbnail Generator	`org.dspace.app.mediafilter.ImageMagickImageThumbnailFilter`	Uses ImageMagick to generate thumbnails for image bitstreams. Requires installation of ImageMagick on your server. See ImageMagick Media Filters.	BMP, GIF, image/png, JPG, TIFF, JPEG, JPEG 2000	no
ImageMagick PDF Thumbnail Generator	`org.dspace.app.mediafilter.ImageMagickPdfThumbnailFilter`	Uses ImageMagick and Ghostscript to generate thumbnails for PDF bitstreams. Requires installation of ImageMagick and Ghostscript on your server. See ImageMagick Media Filters.	Adobe PDF	no

Please note that the `filter-media` script will automatically update the DSpace search index by default.

Enabling/Disabling MediaFilters

The media filter plugin configuration `filter.plugins` in `dspace.cfg` contains a list of all enabled media/format filter plugins (see Configuring Media Filters for more information). By modifying the value of `filter.plugins` you can disable or enable MediaFilter plugins. The `filter.plugins` setting can be set multiple times to enable multiple filters. Each filter must be enabled via its name (see "Name" column in the table above).

```
# Enable the default Text Extractor (for 7.3 or above)
filter.plugins = Text Extractor

# Enable the JPEG thumbnail creator
filter.plugins = JPEG Thumbnail

# Enable the PDF thumbnail creator
filter.plugins = PDFBox JPEG Thumbnail
```

Executing (via Command Line)

The media filter system is intended to be run from the command line (or regularly as a cron task):

```
[dspace]/bin/dspace filter-media
```

With no options, this traverses the asset store, applying media filters to bitstreams, and skipping bitstreams that have already been filtered.

Available Command-Line Options:

- **Help** : `[dspace]/bin/dspace filter-media -h`
  - Display help message describing all command-line options.
- **Force mode** : `[dspace]/bin/dspace filter-media -f`
  - Apply filters to ALL bitstreams, even if they've already been filtered. If they've already been filtered, the previously filtered content is overwritten.
- **Identifier mode** : `[dspace]/bin/dspace filter-media -i 123456789/2`
  - Restrict processing to the community, collection, or item named by the identifier - by default, all bitstreams of all items in the repository are processed. The identifier must be a Handle, not a DB key. This option may be combined with any other option.
- **Maximum mode** : `[dspace]/bin/dspace filter-media -m 1000`
  - Suspend operation after the specified maximum number of items have been processed - by default, no limit exists. This option may be combined with any other option.
- **Plugin mode** : `[dspace]/bin/dspace filter-media -p "PDF Text Extractor","Word Text Extractor"`
  - Apply ONLY the filter plugin(s) listed (separated by commas). By default all named filters listed in the *filter.plugins* field of *dspace.cfg* are applied. This option may be combined with any other option. *WARNING:* multiple plugin names must be separated by a comma (i.e. ',') and NOT a comma followed by a space (i.e. ', ').
- **Skip mode** : `[dspace]/bin/dspace filter-media -s 123456789/9,123456789/100`
  - SKIP the listed identifiers (separated by commas) during processing. The identifiers must be Handles (not DB Keys). They may refer to items, collections or communities which should be skipped. This option may be combined with any other option. *WARNING:* multiple identifiers must be separated by a comma (i.e. ',') and NOT a comma followed by a space (i.e. ', ').
  - NOTE: If you have a large number of identifiers to skip, you may maintain this list, one identifier per line, within a separate file (e.g. *filter-skiplist.txt*). Use the following format to call the program.
    - `[dspace]/bin/dspace filter-media -s $(paste -sd, - < filter-skiplist.txt)`
- **Verbose mode** : `[dspace]/bin/dspace filter-media -v`
  - Print all extracted text and other filter details to STDOUT.

Creating Custom MediaFilters

Adding your own filters is done by creating a class which *implements* the `org.dspace.app.mediafilter.FormatFilter` interface. See the Creating a new Media/Format Filter topic and comments in the source file `FormatFilter.java` for more information. In theory filters could be implemented in any programming language (C, Perl, etc.) However, they need to be invoked by the Java code in the Media Filter class that you create.

Creating a simple Media Filter

New Media Filters **must implement** the *org.dspace.app.mediafilter.FormatFilter* interface. More information on the methods you need to implement is provided in the *FormatFilter.java* source file. For example:

```
public class MySimpleMediaFilter implements FormatFilter
```

Alternatively, you could extend the *org.dspace.app.mediafilter.MediaFilter* class, which just defaults to performing no pre/post-processing of bitstreams before or after filtering.

```
public class MySimpleMediaFilter extends MediaFilter
```

You must give your new filter a "name", by adding it and its name to the *plugin.named.org.dspace.app.mediafilter.FormatFilter* field in *dspace.cfg*. In addition to naming your filter, make sure to specify its input formats in the *filter.<class path>.inputFormats* config item. Note the input formats must match the *short description* field in the Bitstream Format Registry (i.e. *bitstreamformatregistry* table).

```
plugin.named.org.dspace.app.mediafilter.FormatFilter = \
        org.dspace.app.mediafilter.MySimpleMediaFilter = My Simple Text Filter, \ ...

filter.org.dspace.app.mediafilter.MySimpleMediaFilter.inputFormats =
        Text
```

If you neglect to define the *inputFormats* for a particular filter, the *MediaFilterManager* will never call that filter, since it will never find a bitstream which has a format matching that filter's input format(s).

If you have a complex Media Filter class, which actually performs different filtering for different formats (e.g. conversion from Word to PDF **and** conversion from Excel to CSV), you should define this as described in Chapter 13.3.2.2 .

Creating a Dynamic or "Self-Named" Format Filter

If you have a more complex Media/Format Filter, which actually performs **multiple** filtering or conversions for different formats (e.g. conversion from Word to PDF **and** conversion from Excel to CSV), you should have define a class which implements the *FormatFilter* interface, while also extending the Chapter 13.3.2.2 *SelfNamedPlugin* class. For example:

```
public class MyComplexMediaFilter extends SelfNamedPlugin implements FormatFilter
```

Since *SelfNamedPlugins* are self-named (as stated), they must provide the various names the plugin uses by defining a getPluginNames() method. Generally speaking, each "name" the plugin uses should correspond to a different type of filter it implements (e.g. "Word2PDF" and "Excel2CSV" are two good names for a complex media filter which performs both Word to PDF and Excel to CSV conversions).

Self-Named Media/Format Filters are also configured differently in *dspace.cfg*. Below is a general template for a Self Named Filter (defined by an imaginary *MyComplexMediaFilter* class, which can perform both Word to PDF and Excel to CSV conversions):

```
#Add to a list of all Self Named filters
plugin.selfnamed.org.dspace.app.mediafilter.FormatFilter = \
        org.dspace.app.mediafilter.MyComplexMediaFilter
#Define input formats for each "named" plugin this filter implements
        filter.org.dspace.app.mediafilter.MyComplexMediaFilter.Word2PDF.inputFormats = Microsoft Word
        filter.org.dspace.app.mediafilter.MyComplexMediaFilter.Excel2CSV.inputFormats = Microsoft Excel
```

As shown above, each Self-Named Filter class must be listed in the `plugin.selfnamed.org.dspace.app.mediafilter.FormatFilter` item in `dspace.cfg`. In addition, each Self-Named Filter **must** define the input formats for *each named plugin* defined by that filter. In the above example the *MyComplexMediaFilter* class is assumed to have defined two named plugins, `Word2PDF` and `Excel2CSV`. So, these two valid plugin names ("Word2PDF" and "Excel2CSV") **must** be returned by the `getPluginNames()` method of the `MyComplexMediaFilter` class.

These named plugins take different input formats as defined above (see the corresponding *inputFormats* setting).

If you neglect to define the `inputFormats` for a particular named plugin, the `MediaFilterManager` will never call that plugin, since it will never find a bitstream which has a format matching that plugin's input format(s).

For a particular Self-Named Filter, you are also welcome to define additional configuration settings in *dspace.cfg*. To continue with our current example, each of our imaginary plugins actually results in a different output format (Word2PDF creates "Adobe PDF", while Excel2CSV creates "Comma Separated Values"). To allow this complex Media Filter to be even more configurable (especially across institutions, with potential different "Bitstream Format Registries"), you may wish to allow for the output format to be customizable for each named plugin. For example:

```
#Define output formats for each named plugin
filter.org.dspace.app.mediafilter.MyComplexMediaFilter.Word2PDF.output Format = Adobe PDF
filter.org.dspace.app.mediafilter.MyComplexMediaFilter.Excel2CSV.outputFormat = Comma Separated Values
```

Any custom configuration fields in *dspace.cfg* defined by your filter are ignored by the *MediaFilterManager*, so it is up to your custom media filter class to read those configurations and apply them as necessary. For example, you could use the following sample Java code in your *MyComplexMediaFilter* class to read these custom *outputFormat* configurations from *dspace.cfg*:

```
#Get "outputFormat" configuration from dspace.cfg
String outputFormat =  ConfigurationManager.getProperty(MediaFilterManager.FILTER_PREFIX + "." +
MyComplexMediaFilter.class.getName() + "." + this.getPluginInstanceName() + ".outputFormat");
```

Configuration parameters

Property	**textextractor.max-chars** (only in 7.3 or above)
Example Value	textextractor.max-chars = 100000
Informational Note	By default, the "Text Extractor" only extracts the first 100,000 characters of text for full-text indexing. This setting allows you to increase or decrease that default. Set to -1 for no maximum. Keep in mind that larger values (or -1) are more likely to encounter OutOfMemoryException errors when extracting text from very large files. In those scenarios, you may wish to consider instead enabling "textextractor.use-temp-file" below to better control memory usage.
Property	**textextractor.use-temp-file** (only in 7.3 or above)
Example Value	textextractor.use-temp-file = false
Informational Note	By default, the "Text Extractor" will perform all text extraction in memory (i.e. textextractor.use-temp-file=false). This ensures text extraction runs quickly, but it has the risk of hitting OutOfMemoryException errors if you either increase "textextractor.max-chars" or simply don't have much available memory on the server. In those scenarios, you can set "textextractor.use-temp-file=true" in order to tell the text extraction process to extract all text using a temporary file. This *decreases* the memory usage of the text extraction process, but will run slightly slower.
Property	**filter.org.dspace.app.mediafilter.publicPermission**
Example Value	filter.org.dspace.app.mediafilter.publicPermission = JPEGFilter
Informational Note	By default mediafilter derivatives / thumbnails inherit the permissions of the parent bitstream, but you can override this, in case you want to make publicly accessible derivative / thumbnail content, typically the thumbnails of objects for the browse list. List the MediaFilter names that would get public accessible permissions. Any media filters not listed will instead inherit the permissions of the parent bitstream.