Curation System

Curation System

As of release 1.7, DSpace supports running curation tasks, which are described in this section. DSpace 1.7 and subsequent distributions will bundle (include) several useful tasks, but the system also is designed to allow new tasks to be added between releases, both general purpose tasks that come from the community, and locally written and deployed tasks.

Tasks

The goal of the curation system ('CS') is to provide a simple, extensible way to manage routine content operations on a repository. These operations are known to CS as 'tasks', and they can operate on any DSpaceObject (i.e. subclasses of DSpaceObject) - which means Communities, Collections, and Items - viz. core data model objects. Tasks may elect to work on only one type of DSpace object - typically an Item - and in this case they may simply ignore other data types (tasks have the ability to 'skip' objects for any reason). The DSpace core distribution will provide a number of useful tasks, but the system is designed to encourage local extension - tasks can be written for any purpose, and placed in any java package. This gives DSpace sites the ability to customize the behavior of their repository without having to alter - and therefore manage synchronization with - the DSpace source code. What sorts of activities are appropriate for tasks?

Some examples:

- apply a virus scan to item bitstreams (this will be our example below)
- profile a collection based on format types - good for identifying format migrations
- ensure a given set of metadata fields are present in every item, or even that they have particular values
- call a network service to enhance/replace/normalize an item's metadata or content
- ensure all item bitstreams are readable and their checksums agree with the ingest values

Since tasks have access to, and can modify, DSpace content, performing tasks is considered an administrative function to be available only to knowledgeable collection editors, repository administrators, sysadmins, etc. No tasks are exposed in the public interfaces.

Activation

For CS to run a task, the code for the task must of course be included with other deployed code (to `[dspace]/lib`, WAR, etc) but it must also be declared and given a name. This is done via a configuration property in `[dspace]/config/modules/curate.cfg` as follows:

```
plugin.named.org.dspace.curate.CurationTask = \
org.dspace.curate.ProfileFormats = profileformats, \
org.dspace.curate.RequiredMetadata = requiredmetadata, \
org.dspace.curate.ClamScan = vscan
```

For each activated task, a key-value pair is added. The key is the fully qualified class name and the value is the *taskname* used elsewhere to configure the use of the task, as will be seen below. Note that the curate.cfg configuration file, while in the config directory, is located under 'modules'. The intent is that tasks, as well as any configuration they require, will be optional 'add-ons' to the basic system configuration. Adding or removing tasks has no impact on dspace.cfg.

For many tasks, this activation configuration is all that will be required to use it. But for others, the task needs specific configuration itself. A concrete example is described below, but note that these task-specific configuration property files also reside in `[dspace]/config/modules`

Writing your own tasks

A task is just a java class that can contain arbitrary code, but it must have 2 properties:

First, it must provide a no argument constructor, so it can be loaded by the PluginManager. Thus, all tasks are 'named' plugins, with the taskname being the plugin name.

Second, it must implement the interface 'org.dspace.curate.CurationTask'

The CurationTask interface is almost a 'tagging' interface, and only requires a few very high-level methods be implemented. The most significant is:

```
int perform(DSpaceObject dso);
```

The return value should be a code describing one of 4 conditions:

- 0 : SUCCESS the task completed successfully
- 1 : FAIL the task failed (it is up to the task to decide what 'counts' as failure - an example might be that the virus scan finds an infected file)
- 2 : SKIPPED the task could not be performed on the object, perhaps because it was not applicable
- -1 : ERROR the task could not be completed due to an error

If a task extends the AbstractCurationTask class, that is the only method it needs to define.

Task Invocation

Tasks are invoked using CS framework classes that manage a few details (to be described below), and this invocation can occur wherever needed, but CS offers great versatility 'out of the box':

On the command line

A simple tool 'CurationCli' provides access to CS via the command line. This tool bears the name 'curate' in the DSpace launcher. For example, to perform a virus check on collection '4':

```
[dspace]/bin/dspace curate -t vscan -i 123456789/4
```

The complete list of arguments:

```
-t taskname: name of task to perform
-T filename: name of file containing list of tasknames
-e epersonID: (email address) will be superuser if unspecified
-i identifier: Id of object to curate. May be (1) a handle (2) a workflow Id or (3) 'all' to operate on the
whole repository
-q queue: name of queue to process - -i and -q are mutually exclusive
-v emit verbose output
-r - emit reporting to standard out
```

As with other command-line tools, these invocations could be placed in a cron table and run on a fixed schedule, or run on demand by an administrator.

In the admin UI

In the XMLUI, there is a 'Curate' tab (appearing within the 'Edit Community/Collection/Item') that exposes a drop-down list of configured tasks, with a button to 'perform' the task, or queue it for later operation (see section below). Not all activated tasks need appear in the Curate tab - you filter them by means of a configuration property. This property also permits you to assign to the task a more user-friendly name than the PluginManager *taskname*. The property resides in `[dspace]/config/modules/curate.cfg`:

```
ui.tasknames = \
     profileformats = Profile Bitstream Formats, \
     requiredmetadata = Check for Required Metadata
```

When a task is selected from the drop-down list and performed, the tab displays both a phrase interpreting the 'status code' of the task execution, and the 'result' message if any has been defined. When the task has been queued, an acknowledgement appears instead. You may configure the words used for status codes in curate.cfg (for clarity, language localization, etc):

```
ui.statusmessages = \
    -3 = Unknown Task, \
    -2 = No Status Set, \
    -1 = Error, \
     0 = Success, \
     1 = Fail, \
     2 = Skip, \
     other = Invalid Status
```

In workflow

CS provides the ability to attach any number of tasks to standard DSpace workflows. Using a configuration file `[dspace]/config/workflow-curation.xml`, you can declaratively (without coding) wire tasks to any step in a workflow. An example:

```
<taskset-map>
   <mapping collection-handle="default" taskset="cautious" />
</taskset-map>
<tasksets>
  <taskset name="cautious">
    <flowstep name="step1">
      <task name="vscan">
        <workflow>reject</workflow>
        <notify on="fail">$flowgroup</notify>
        <notify on="fail">$colladmin</notify>
        <notify on="error">$siteadmin</notify>
      </task>
    </flowstep>
  </taskset>
</tasksets>
```

This markup would cause a virus scan to occur during step one of workflow for any collection, and automatically reject any submissions with infected files. It would further notify (via email) both the reviewers (step 1 group), and the collection administrators, if either of these are defined. If it could not perform the scan, the site administrator would be notified.

The notifications use the same procedures that other workflow notifications do - namely email. There is a new email template defined for curation task use: `[dspace]/config/emails/flowtask_notify`. This may be language-localized or otherwise modified like any other email template.

Like configurable submission, you can assign these task rules per collection, as well as having a default for any collection.

In arbitrary user code

If these pre-defined ways are not sufficient, you can of course manage curation directly in your code. You would use the CS helper classes. For example:

```
Collection coll = (Collection)HandleManager.resolveToObject(context, "123456789/4");
Curator curator = new Curator();
curator.addTask("vscan").curate(coll);
System.out.println("Result: " + curator.getResult("vscan"));
```

would do approximately what the command line invocation did. the method 'curate' just performs all the tasks configured
(you can add multiple tasks to a curator).

Asynchronous (Deferred) Operation

Because some tasks may consume a fair amount of time, it may not be desirable to run them in an interactive context. CS provides a simple API and means to defer task execution, by a queuing system. Thus, using the previous example:

```
    Curator curator = new Curator();
    curator.addTask("vscan").queue(context, "monthly", "123456789/4");
```

would place a request on a named queue "monthly" to virus scan the collection. To read (and process) the queue, we could for example:

```
[dspace]/bin/dspace curate -q monthly
```

use the command-line tool, but we could also read the queue programmatically. Any number of queues can be defined and used as needed.
In the administrative UI curation 'widget', there is the ability to both perform a task, but also place it on a queue for later processing.

Task Output and Reporting

Few assumptions are made by CS about what the 'outcome' of a task may be (if any) - it. could e.g. produce a report to a temporary file, it could modify DSpace content silently, etc But the CS runtime does provide a few pieces of information whenever a task is performed:

Status Code

This was mentioned above. This is returned to CS whenever a task is called. The complete list of values:

```
-3 NOTASK - CS could not find the requested task
-2 UNSET  - task did not return a status code because it has not yet run
-1 ERROR - task could not be performed
 0 SUCCESS - task performed successfully
 1 FAIL - task performed, but failed
 2 SKIP - task not performed due to object not being eligible
```

In the administrative UI, this code is translated into the word or phrase configured by the *ui.statusmessages* property (discussed above) for display.

Result String

The task may define a string indicating details of the outcome. This result is displayed, in the 'curation widget' described above:

```
"Virus 12312 detected on Bitstream 4 of 1234567789/3"
```

CS does not interpret or assign result strings, the task does it. A task may not assign a result, but the 'best practice' for tasks is to assign one whenever possible.

Reporting Stream

For very fine-grained information, a task may write to a *reporting* stream. This stream is sent to standard out, so is only available when running a task from the command line. Unlike the result string, there is no limit to the amount of data that may be pushed to this stream.

The status code, and the result string are accessed (or set) by methods on the Curation object:

```
Curator curator = new Curator();
curator.addTask("vscan").curate(coll);
int status = curator.getStatus("vscan");
String result - curator.getResult("vscan");
```

Task Annotations

CS looks for, and will use, certain java annotations in the task Class definition that can help it invoke tasks more intelligently. An example may explain best. Since tasks operate on DSOs that can either be simple (Items) or containers (Collections, and Communities), there is a fundamental problem or ambiguity in how a task is invoked: if the DSO is a collection, should the CS invoke the task on each member of the collection, or does the task 'know' how to do that itself? The decision is made by looking for the @Distributive annotation: if present, CS assumes that the task will manage the details, otherwise CS will walk the collection, and invoke the task on each member. The java class would be defined:

```
@Distributive
public class MyTask implements CurationTask
```

A related issue concerns how non-distributive tasks report their status and results: the status will normally reflect only the last invocation of the task in the container, so important outcomes could be lost. If a task declares itself @Suspendable, however, the CS will cease processing when it encounters a FAIL status. When used in the UI, for example, this would mean that if our virus scan is running over a collection, it would stop and return status (and result) to the scene on the first infected item it encounters. You can even tune @Supendable tasks more precisely by annotating what invocations you want to suspend on. For example:

```
@Suspendable(invoked=Curator.Invoked.INTERACTIVE)
public class MyTask implements CurationTask
```

would mean that the task would suspend if invoked in the UI, but would run to completion if run on the command-line.

Only a few annotation types have been defined so far, but as the number of tasks grow, we can look for common behavior that can be signaled by annotation. For example, there is a @Mutative type: that tells CS that the task may alter (mutate) the object it is working on.

Starter Tasks

DSpace 1.7 bundles a few tasks and activates two (2) by default to demonstrate the use of the curation system. These may be removed (deactivated by means of configuration) if desired without affecting system integrity. Each task is briefly described here.

Bitstream Format Profiler

The task with the taskname 'formatprofiler' (in the admin UI it is labeled "Profile Bitstream Formats") examines all the bitstreams in an item and produces a table ("profile") which is assigned to the result string. It is activated by default, and is configured to display in the administrative UI. The result string has the layout:

```
    10 (K) Portable Network Graphics
    5  (S) Plain Text
```

where the left column is the count of bitstreams of the named format and the letter in parentheses is an abbreviation of the repository-assigned support level for that format:

```
   U  Unsupported
   K  Known
   S  Supported
```

The profiler will operate on any DSpace object. If the object is an item, then only that item's bitstreams are profiled; if a collection, all the bitstreams of all the items; if a community, all the items of all the collections of the community.

Required Metadata

The 'requiredmetadata' task examines item metadata and determines whether fields that the web submission (input-forms.xml) marks as required are present. It sets the result string to indicate either that all required fields are present, or constructs a list of metadata elements that are required but missing. When the task is performed on an item, it will display the result for that item. When performed on a collection or community, the task be performed on each item, and will display the *last* item result. If all items in the community or collection have all required fields, that will be the last in the collection. If the task fails for any item (i.e. the item lacks all required fields), the process is halted. This way the results for the 'failed' items are not lost.

Virus Scan

The 'vscan' task performs a virus scan on the bitstreams of items using the ClamAV software product.
Clam AntiVirus is an open source (GPL) anti-virus toolkit for UNIX. A port for Windows is also available. The virus scanning curation task interacts with the ClamAV virus scanning service to scan the bitstreams contained in items, reporting on infection(s). Like other curation tasks, it can be run against a container or item, in the GUI or from the command line. It should be installed according to the documentation at http://www.clamav.net. It should not be installed in the dspace installation directory. You may install it on the same machine as your dspace installation, or on another machine which has been configured properly.

Setup the service from the ClamAV documentation.

This plugin requires a ClamAV daemon installed and configured for TCP sockets. Instructions for installing ClamAV (http://www.clamav.net/doc/latest/**clam doc**.pdf)

NOTICE: The following directions assume there is a properly installed and configured clamav daemon. Refer to links above for more information about ClamAV.
The Clam anti-virus database must be updated regularly to maintain the most current level of anti-virus protection. Please refer to the ClamAV documentation for instructions about maintaining the anti-virus database.

DSpace Configuration

In `[dspace]/config/modules/curate.cfg`, activate the task:

- Add the plugin to the comma separated list of curation tasks.

```
### Task Class implementations
plugin.named.org.dspace.curate.CurationTask = \
org.dspace.curate.ProfileFormats = profileformats, \
org.dspace.curate.RequiredMetadata = requiredmetadata, \
org.dspace.curate.ClamScan = vscan
```

- Optionally, add the vscan friendly name to the configuration to enable it in the administrative it in the administrative user interface.

```
ui.tasknames = \
profileformats = Profile Bitstream Formats, \
requiredmetadata = Check for Required Metadata, \
vscan = Scan for Viruses
```

In `[dspace]/config/modules`, edit configuration file clamav.cfg:

```
service.host = 127.0.0.1
Change if not running on the same host as your DSpace installation.
service.port = 3310
Change if not using standard ClamAV port
socket.timeout = 120
Change if longer timeout needed
scan.failfast = false
Change only if items have large numbers of bitstreams
```

Task Operation from the GUI

Curation tasks can be run against container and item dspace objects by e-persons with administrative privileges. A curation tab will appear in the administrative ui after logging into DSpace:

1. Click on the curation tab.
2. Select the option configured in ui.tasknames above.
3. Select Perform.

Task Operation from the curation command line client

To output the results to the console:

```
[dspace]/bin/dspace curate -t vscan -i <handle of container or item dso> -r -
```

Or capture the results in a file:

```
[dspace]/bin/dspace curate -t vscan -i <handle of container or item dso> -r - > /<path...>/<name>
```

Table 1 – Virus Scan Results Table

GUI (Interactive Mode)	FailFast	Expectation
Container	T	Stop on 1st Infected Bitstream
Container	F	Stop on 1st Infected Item
Item	T	Stop on 1st Infected Bitstream
Item	F	Scan all bitstreams

Command Line		
Container	T	Report on 1st infected bitstream within an item/Scan all contained Items
Container	F	Report on all infected bitstreams/Scan all contained Items
Item		Report on 1st infected bitstream
Item		Report on all infected bitstreams