**Setting up and testing the harvester**

Preliminary steps:

* Set up DSpace and deploy either XMLUI or JSPUI
* Obtain the address of the OAI-PMH data provider you wish to harvest from (or simply use the example provided here)

**XMLUI**

Setting up a collection (Collection Edit screen):

* Login and create a new collection
* Go to the tab named “Content Source” that now appears next to “Edit Metadata” and “Assign Roles” in the collection edit screen.
* The two counter source options are standard (selected by default) and harvested. Selected “harvests from external source” and click Save.
* A new set of menus appears to configure the harvesting settings:
  + “OAI Provider” is the URL of the OAI-PMH provider that the content from this collection should be harvested from. The PMH provider deployed with DSpace typically has the form: “http://dspace.url/oai/request”. For this example, use “http://web01.library.tamu.edu/oai-h151/request”
  + “OAI Set id” is the setSpec of the collection you wish to harvest from. Use “hdl\_1969.1\_5671” for this example.
  + “Metadata Format” determines the format that the descriptive metadata will be harvested in. Since DSpace stores metadata in its own internal format, not all metadata values might get harvested if a specific format is specified. Select “DSpace Intermediate Metadata” if available and “Simple Dublin Core” otherwise.
  + Clicking the Test Settings button will verify the settings supplied in the previous steps and will usually let you know what, if anything is missing or does not match up.
* The list of radio buttons labeled “Content being harvested” allows you to select the harvest level. The first one requires no OAI-ORE support on the part of the provider and can be used to harvest metadata from any provider compliant with the OAI-PMH 2.0 specifications. The middle options will harvest the metadata and generate links to bitstreams stored remotely, while the last one will do perform full local replication.
* Select the middle option and click Save.

At this point the settings are saved and the menu changes to provide three options:

* “Change Settings” takes you back to the edit screen
* “Import Now” performs a single harvest from the remote collection into the local one. Success, notes, and errors encountered in the process will be reflected in the “Last Harvest Result” entry. More detailed information is available in the dspace log. Note that the whole harvest cycle is executed within a single HTTP request and will time out for large collections. For this reason, it is advisable to use the automatic harvest scheduler set up either in XMLUI or from command line. If the scheduler is running, “Import Now” will handle the harvest task as a separate thread.
* “Reset and Reimport Collection” will perform the same function as “Import Now”, but will clear the collection of all existing items before doing so.

Setting up automatic harvesting (Control Panel screen):

* A new tab, Harvesting, has been added under Administrative -> Control Panel
* The panel offers the following information:
  + Available actions
    - Start Harvester: starts the scheduler. From this point on, all properly configured collections (listed on the next line) will be harvested at regular intervals. This interval can be changed in dspace.cfg using the “harvester.harvestFrequency” parameter.
    - Pause: the “nice” stop; waits for the active harvests to finish, saves the state/progress and pauses execution. Can be either resumed or stopped.
    - Stop: the “full stop”; waits for the current item to finish harvesting, and aborts further execution.
    - Reset Harvest Status: since stopping in the middle of a harvest is likely to result in collections getting “stuck” in the queue, the button is available to clear all states.

**JSPUI**

Setting up a collection (Collection Edit screen):

* Login and create a new collection
* After the collection creation steps, a new set of menus appears at the bottom of the Edit Collection screen to configure the harvesting settings:
  + “Content Source” determines whether the collecton is a standard DSpace collection (default) or one that is harvested from an external source. Select the second option.
  + “OAI Provider” is the URL of the OAI-PMH provider that the content from this collection should be harvested from. The PMH provider deployed with DSpace typically has the form: “http://dspace.url/oai/request”. For this example, use “http://web01.library.tamu.edu/oai-h151/request”
  + “OAI Set id” is the setSpec of the collection you wish to harvest from. Use “hdl\_1969.1\_5671” for this example.
  + “Metadata Format” determines the format that the descriptive metadata will be harvested in. Since DSpace stores metadata in its own internal format, not all metadata values might get harvested if a specific format is specified. Select “DSpace Intermediate Metadata” if available and “Simple Dublin Core” otherwise.
  + The list of radio buttons labeled “Content being harvested” allows you to select the harvest level. The first one requires no OAI-ORE support on the part of the provider and can be used to harvest metadata from any provider compliant with the OAI-PMH 2.0 specifications. The middle options will harvest the metadata and generate links to bitstreams stored remotely, while the last one will do perform full local replication.
* Click Update to save the settings

At this time, JSPUI does not provide the ability to launch harvest tasks from the interface. The command-line utility should be used to harvest individual collections and to start automatic scheduler.

**Command-line utility**

The command-line harvesting utility offers all the features available in XMLUI and JSPUI and a few others. The only caveat is that a collection must already have been created through one of the other interfaces before it can be configured for harvesting through the command-line tool. It is executed using dsrun (use –h flag to get a list of available parameters):

[dspace-insall]/bin/dsrun org.dspace.app.harvest.Harvest