Running a simple harvest

Walkthrough of the definition and execution of a simple harvest.

- Running a simple harvest
  - Setting up the harvest
  - Viewing the results

Running a simple harvest

The system is now up and running, and you can try out the harvesting and archiving capabilities.

This section will guide you through the steps needed to

- harvest and store a domain
- browse the harvested material in a browser

Setting up the harvest

Start the program as described in section "Starting simple_harvest version".

Open http://localhost:8074/HarvestDefinition in a browser on the local machine. (Replace the host name if the QUICKSTART is running on another machine)

You can now define a new harvest.

Click 'Selective Harvests' under menu 'Definitions'

- Selective Harvests
  - Show inactive harvest definitions
  - No selective harvests defined
  - Create new selective harvest definition

Click 'Create new selective harvest definition' under the (empty) table of existing harvests.

Enter a name for the harvest in the top. Enter some second-level domain name (e.g., netarkivet.dk) in the box and press 'Add domains'. Assuming the domain doesn't already exist in the database, the system suggests you add it. Click 'Create and add to harvest definition'.

Preferably the domain should be one that you know you have permission to harvest. By default, NetarchiveSuite will harvest up to 1GB of data from a domain so you may wish to choose a small domain for your first tests. You can add more domains if you want by repeating the procedure, but in this example we will only add one domain.
You can now click 'Save' on the 'Selective Harvest' page.

Click 'Activate' for the newly defined harvest. NetarchiveSuite will now generate harvest jobs for the harvest definition.
Go to the Job Status page by clicking 'Harvest status'. Set wanted jobs status to 'All' and click 'Show'. Refresh the page periodically until a job appears and changes to state "Started". This should take no more than two minutes. At this point, a harvester has started harvesting, using the Heritrix web harvester.

Now you can monitor the system state for what is going on in the various components. That way you can see how the harvester is proceeding with the job:

Go to the System Status page by clicking 'Systemstate'. Click on the application HarvestControllerApplication. The most recent log record will give status information from Heritrix. You can find more application information by clicking on 'Show all' in the Index column.
You can find more details about the running job by going to the Running Jobs page:

<table>
<thead>
<tr>
<th>Machine (hide)</th>
<th>Application (show all, hide)</th>
<th>Channel (hide)</th>
<th>Use Replica (hide)</th>
<th>Index (show all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kb-test-har-003</td>
<td>HarvestControllerApplication</td>
<td>LOWPRIORITY</td>
<td>KBN</td>
<td>0</td>
</tr>
<tr>
<td>kb-test-har-004</td>
<td>HarvestControllerApplication</td>
<td>HIGHPRIORITY</td>
<td>KBN</td>
<td>0</td>
</tr>
<tr>
<td>kb-test-har-004</td>
<td>HarvestControllerApplication</td>
<td>LOWPRIORITY</td>
<td>KBN</td>
<td>0</td>
</tr>
<tr>
<td>sb-test-har-001</td>
<td>HarvestControllerApplication</td>
<td>HIGHPRIORITY</td>
<td>SBN</td>
<td>0</td>
</tr>
</tbody>
</table>

The job will appear as running and the progress of the jobs can be followed here. A link to the running Heritrix3 crawler can be found in the Host column, as long as the job is running. The Heritrix 3 console can be accessed using the standard Heritrix login "admin" and Password "adminPassword" (Note: you will need to add the name of your PC as an exception to your browser’s proxy configuration. Clicking on the JobID will take you to a page where you can view information on the state of the job and control most of the functionality of the heritrix harvesting process directly from NetarchiveSuite:
By clicking on the various buttons, you can see what the job has already harvested (the crawl-log), see and manipulate what the job is going to harvest (the frontier), monitor the output of the job, pause, unpause, and if necessary, terminate the job.

Now go to the Job status page by clicking ‘Harvest status’. Set wanted jobs status to ‘All’ and click ‘Show’. Refresh the page until your job changes state to “Done”.

**Viewing the results**

Harvested jobs can be viewed in an ordinary browser. Part of the NetarchiveSuite is a "viewerproxy", that integrates with your browser to show you harvested material for Quality Assurance.

In order to use viewerproxy it is essential that you have followed the instructions for proxy setup. Once a harvest has completed, you can use the viewerproxy part to view them. Before it is ready, it needs to know which material you wish to browse.

1. Go to the ‘Harvest Status’ page, select to show ‘All’ jobs and click ‘Show’. Click on the link with the Job id.
2. Click on ‘Select this job for QA with viewerproxy’. This will make the viewerproxy browse in this job. It will take it a while to generate an index. It will then go to the viewerproxy status page.

Generating index ‘Job 17’ for jobs [17]
Now simply enter (in the browser url field) the URL that you started harvesting from (with www), e.g. www.netarkivet.dk. It shows you the harvested material. If you go to a URL in another domain, you will get an error. Depending on the layout of the domain you harvested, there may also be missing pages or images from that domain.

The NetarchiveSuite allows automatic collection of unharvested URLs during browsing, i.e. the NetarchiveSuite allows you to browse in the collected material while it automatically collects URLs for missing pages or images that you request. This makes it easy to identify missing harvested material during Quality Assurance of the harvested material.

To try this, go back to the Viewerproxy status page and click 'Start collecting URLs'. Now browse in the collected material until you find a page or image that did not get harvested. Go back to the Viewerproxy status page and click 'Show collected URLs'.

The list will contain several URLs, including the ones you just requested and found missing during collection of URLs.