Harvest Status

All jobs

Harvest Status (or All Jobs) in the left menu by default shows a list of all jobs in the system, ranked in chronological order, most recent first (in NetarchiveSuite 5.2 and above - this is a change from previous versions).

The top part of the page has options to allow you to filter or select jobs according to
- Status
- Harvest name
- Job number (ID)
- Start and end date

and also allows you to control how many results to display and their sort order.

For each job the page shows information about the job and its status as well as information about errors (harvest errors or upload errors) and number of configurations in the job.

- Click Job ID if you want to check details on a specific job. See Job list links.
- Click Harvest name if you want to check details on the history of a specific harvest definition. See History of a harvest definition for details.
- Click Run number if you want to check details on a specific run of that harvestdefinition – note that a run can consist of multiple jobs. See Specific run listing.
In case of Harvest errors, a Restart button will appear in the Harvest errors column, and the operator can choose to resubmit that specific job to be harvested again. When resubmitting a failed job, the status will say 'Resubmitted' and a link will appear pointing to the new Job.

**Job list links**

Clicking on a jobID on any of the harvest history pages will give you a detailed report on the job.

### Details for Job 13

<table>
<thead>
<tr>
<th>Job ID</th>
<th>Type</th>
<th>Harvest name</th>
<th>Run number</th>
<th>Created</th>
<th>Start time</th>
<th>End time</th>
<th>Status</th>
<th>Harvest errors</th>
<th>Upload errors</th>
<th>Object limit</th>
<th>Byte limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>LOWPRODITY</td>
<td>snap</td>
<td>1</td>
<td>2016/11/07 00:44:48</td>
<td>2016/11/07 00:47:42</td>
<td>2016/11/07 00:49:51</td>
<td>Done</td>
<td>0</td>
<td>10</td>
<td>100000</td>
<td></td>
</tr>
</tbody>
</table>

### QA job selection

Select this job for QA with viewproxy.

The link above will select this job for the viewproxy browse index. This will only work if your browser is set up to use the viewproxy as web proxy.

### Included domains and configurations

<table>
<thead>
<tr>
<th>Domain</th>
<th>Configuration</th>
<th>Bytes Harvested</th>
<th>Documents Harvested</th>
<th>Stopped due to</th>
</tr>
</thead>
<tbody>
<tr>
<td>statsbiblioteket.dk</td>
<td>defaultconfig</td>
<td>0</td>
<td>5</td>
<td>Domain-config byte limit reached</td>
</tr>
</tbody>
</table>

### Seed list

www.statsbiblioteket.dk

### Harvest information for job

Browse reports for jobs
Browse harvest files for job

Display all crawl.log lines matching this regexp
Browse only relevant crawl.log lines for this domain statsbiblioteket.dk

### Harvest order template (based on default_orderxml)

Show harvest template for job 13 (text/plain)

This page gives all the information available about the job itself (e.g. max-bytes limit) and about the individual domains included in the job.

Furthermore the page shows the complete seedlist used with the job and the complete Harvest crawler-bean template as well as detailed error information in case of errors. The latter are mainly for advanced users debugging specific crawls where things didn’t go as expected.

### Harvest definition job history

Searching for 'snap' returned 1 hits:

Search results: 1, displaying results 1 to 1.

### Harvest history for full harvest snap

<table>
<thead>
<tr>
<th>Run number</th>
<th>Start time</th>
<th>End time</th>
<th>Bytes Harvested</th>
<th>Documents Harvested</th>
<th>Total number of jobs</th>
<th>Number of failed jobs</th>
<th>Number of resubmitted jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2016/11/07 00:44:51</td>
<td>2016/11/07 00:56:42</td>
<td>0</td>
<td>122</td>
<td>4 Show jobs</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The history page for a harvestdefinition is the same as you can reach from the frontpage with the History buttons. This history page gives you further information for each run of the harvestdefinition: Start time, End time, number of bytes harvested and number of documents harvested. The page also show how many jobs each run consists of and how many of these failed and eventually got resubmitted.

### Specific run listing
If a link in the 'Run number' column is clicked, the jobs for the specific run of the harvests definition is listed. A scheduled harvest definition might generate a number of jobs depending on the configurations and number of domains. See Harvester design for details on how jobs are generated.

Harvester design
Details on a terminated job
If you terminate a running job in the Heritrix GUI, you can view Job Details and see that the job is stopped due to "Harvesting aborted". If the job was part of a larger harvest, that includes several jobs, all the finished jobs, will appear as "Done". Only the ones that were actually stopped will appear as stopped due to "Harvesting aborted" for some domains.

All job per domain
This link is obsolete and will just advise the user to access the harvest history for a domain functionality from the domain details page.

Running jobs
The running jobs page displays details about the jobs currently being run by the harvesters. The information shown here is continuously extracted from the running Heritrix instance and returned to the GUI.

All Running Jobs
There is 1 running job.

Under "host" there will appear, whenever it is available, a link to the Heritrix GUI. In a change to earlier versions on NAS, the Job ID on this page is now a link to the Heritrix Remote Access page for the job not, as previously, a link to the Job definition in the NAS system.