TEST1

Test basic start and stop of selective-,event- and snapshot harvesting, scheduling and deduplication.

1. Check of monitoring and basic settings

Check that

- No error messages are received at start up
- The monitoring works
- The database works and contains test data

Do following in a browser:

Start Program

1. Go to http://$GUIadminserver:$http-port/HarvestDefinition/ http://$GUIadminserver:$http-port/HarvestDefinition/ where GUIadminserver and http-port are specified in the deploy configuration file under the application named dk.netarkivet.common.webinterface.GUIApplication. In the one-machine setup (deploy_example_one_machine.xml) the link will be : http://localhost:8074

Check that JMX works (Partially tested by *SystemOverviewTest#generalTest*)

- Click 'Systemstate'->'Overview of the system state' (At first, only the 'SystemState' link is visible, which leads to the same script, viz. Status /Monitor-JMXsummary.jsp
- Check that all internally developed applications are up and running. This depends on the configuration, of course, but there should be
    - One GUIApplication
    - One HarvestJobManagerAppplication
    - One or more HarvestControllerServer. One of these should be HIGHPRIORITY, if you need to run selective/event harvests, and/or LOWPRIORITY, if you need to run snapshot harvests
    - One IndexServerApplication
    - One or more ViewerProxy applications
    - One `WaybackIndexerApplication (optional)`
    - `One` AggregatorApplication (Optional)
    - If you're using the distributed archive solution (ie using the JMSArcrepositoryClient and not the LocalArcRepositoryClient) there should also be
        - One ArcRepositoryApplication
        - One or more BitarchiveServer applications per bitarchive replica in your configuration
        - One ChecksumFileApplication representing a checksum replica (if you have a checksum replica)
        - One BitarchiveMonitorServer for each bitarchive replica in your configuration
- Check that last status message for each application do not contain errors or warnings
- Check that there are no empty log messages
- Click on an physical location in the 'Location' column e.g. "K" (you might need add "location" to the shown columns by clicking on "Location" in the "show:" line just above the column headers.
- Check that you now only see relevant SW applications for the chosen location
- Click 'Show all' in the 'Location' header
- Check that you return to the full listening again
- Repeat the above 4 steps for "Machine", "HTTP Port", "Application", "Instance Id", "Priority", "Use Replica" and "Index" (Index shows all log lines in the given appl. log)

The Instance ID column in the System overview GUI is a technical suffix to the Application column to separate more than one Application of the same type on the same server. If there is only 1 Application on a server it will normally be empty. If there is more than 1 Application of the same type on the same server, there must be added a suffix i.e. an Instance ID. It is userdefined (in the deploy script) and must be unique.

- Click 'Systemstate' -> 'Overview of the system state'
- Check that you are back to the full overview with log line 0

Check that basic database data is present

- Click 'Definitions'->'Find Domain(s)'
- Search for **netarkivet.dk** by writing this text and click 'Search'
- Check that the GUI returns a result-set of one, namely the domain **netarkivet.dk**
- Click on the link **netarkivet.dk**, and the page for domain **netarkivet.dk** should be shown without errors
- Click 'Edit' on the line configuration line for defaultconfig
- Check that Name is "defaultconfig"
- Check that Harvest template is "default_orderxml"
- Check that Maximum number of objects is "2,000" (in some languages (e.g. Danish) this is represented as 2.000
- Check that Maximum number of bytes is "500,000,000" (in some languages (e.g. Danish) this is represented as 500.000.000
- Check that max hops is set to 20
- Check that Honour robots.txt? is unchecked
- Check that Extract Javascript? is checked

Check that mail recipients specified in the start of this test receive no error mails

- Check that there are no mails with error messages about non-existing files
- Check that there are no mails with error messages about applications that could not be started

Check that the default_orderxml template is a Heritrix3 template(go to the **Edit Harvest Templates** tab, and retrieve the default_orderxml: If the file-header contains "HERITRIX 3 CRAWL JOB CONFIGURATION FILE ", it is ok.

Check that the DispositionChain includes a deduplicator:

```
<ref bean="DeDuplicator"/>
```

2. Running selective harvest

Partially tested by *SelectiveHarvestTest*

1. Go to http://$GUIadminserver:$http-port/HarvestDefinition/where GUIadminserver and http-port are specified in the deploy configuration file under the application named dk.netarkivet.common.webinterface.GUIApplication. In the one-machine setup (deploy_example_one_machine.xml) the link will be : http://localhost:8074
2. Make a new selective harvest definition with a name you can remember
   a. Click 'Definitions'->'Selective Harvests' in the left menu
   b. *Click 'Create new harvest definition' in the bottom of the main window
   c. Fill in the Harvest name and note the name for later use (from now referred as **sh.name**)
   d. Choose "Once_a_week" in the drop down list for 'Schedule'
   e. In the 'Enter Domain...' window add the name of a domain not already in the system (e.g. mazda.dk) and click 'Add domains'
   f. There should be a button "Create and add to the harvest definition" shown. Click on it.
   g. Click 'Save'
3. Activate the selective harvest
   a. Click 'Activate' in column 5 on the line with the **sh. name**
   b. Check that the time in the "Next Run" column time on the line with the **sh. name** is now.
4. Check harvest status of the selective harvest
   a. Click 'Harvest status'->'All Jobs' in the left menu
   b. Click the "Show" button, until the name appears in a new job line (approx. after a minute)
   c. Check that the job has status "NEW", it may have turned into status "SUBMITTED" or status "STARTED" before you see it.
5. Check job creation in the system status for the selective harvest
   a. Click 'Systemstate'->'Overview of the system state'
   b. Find and click 'HarvestJobManagerApplication' in the 'Application' column.
   c. Click 'show all' in the header.
   d. Check that there exists a line with the message "INFO: Created 1 jobs for harvest definition ' and a line after that "INFO: Job #1 submitted, and later the line: "INFO: Job #1 has been started by the harvester."

2.1 Run an Umbra Harvest

1. For the same domain as above create a new Harvest Configuration using template default_orderxml (the other template available has not been modified for use with Umbra)
2. Create a Selective Harvest definition using this new configuration
3. Use the Mapping functionalityunder Harvest Channels in the GUI to map this new configuration to the UMBRA channel
4. Activate the harvest and wait for it to complete
5. Check the crawl log for the completed harvest for the strings "sentToAMQP" and "receivedFromAMQP"

3. Define and run an event harvest

This page describes how to define and run an event harvest. It also test that seeds lists are created from first and second level definition of the domain names.

1. Make a new selective (event) harvest definition with a name you can remember
    a. Click 'Definitions'->'Selective Harvests' in the left menu
    b. Click 'Create new harvestdefinition' in the bottom of the main window
    c. Fill in the Harvest name and note the name for later use (from now referred as **EH**)
    d. Choose "'Once_an_hour'" in the drop down list for 'Schedule'
    e. Click Save (DO NOT CLICK ACTIVATE YET)
2. Add seeds to the selective (event) harvest
    a. Click 'Edit' in column 6 on the line with the **EH**
    b. Write domain list from 'Seed list 1' given below to a file on your desktop e.g. notepad)
    c. Click 'Add seeds from a file' at the bottom of the main page
    d. Click 'Browse" and pick up the just created file with seeds
    e. Choose **default_orderxml** in the drop-down list for 'Harvest template' (set **maxobjects pr domain** to 500; **max bytes** to 400.000.000, **maxhops** to 0, **obey robots.txt?** unchecked and **extract_javascript** checked) [previously used template frontpages]
    f. Click 'Insert'
    g. Now click 'Add seeds'
    h. Choose **default_orderxml** in the drop-down list for 'Harvest template'
    i. Write domain list from 'Seed list 2' given below (you can cut and paste from this page) (set **maxobjects pr domain** to 300; **max bytes** to 500.000.000, **maxhops** to 2, **obey robots.txt?** unchecked and **extract_javascript** checked) [previously used template frontpages_2levels]
    j. Click 'Insert'
    k. *Click 'Save'
3. Check that seed lists for domains in Seed list 1 has changed correspondingly (You have to click on Show unused configurations/seedlists show all)
    a. For each of the domains **raeder.dk**, **netarkivet.dk** do:
    b. Click 'Definitions'->'Find Domain(s)'
    c. Search for domain by writing its name as text and click 'Search'
    d. Check that there exists a configuration with the name "**EH_default_orderxml_400000000Bytes_500Objects**" (verify that the config has maxHops=0, obey robots unchecked, extract javascript checked)
    e. Check that there exists a seed list with the name "**EH_default_orderxml_400000000Bytes_500Objects**
    f. Click 'Edit' in the line with seed list "**EH_default_orderxml_400000000Bytes_500Objects**
    g. Check that the seed list shown corresponds to the seed list for the domain (see below)
    h. Check that seed lists for domains in Seed list 2 has changed correspondingly (you have to click on Show unused configurations/seedlists show all)
    i. For the domains **kaarefc.dk**, **netarkivet.dk** do:
    j. Click 'Definitions'->'Find Domain(s)'
    k. Search for the domain by writing this text (either kaarefc.dk or netarkivet.dk) and click *Search*
    l. Check that there exists a configuration with name **EH_default_orderxml_500000000Bytes_300Objects** (verify that the config has maxHops=2)
    m. Check that there exists a seed list with the name **EH_default_orderxml_500000000Bytes_300Objects**
    n. Click 'Edit' in the line with seed list **EH_default_orderxml_500000000Bytes_300Objects**
    o. Check that the seed list shown corresponds to the seed list for the domain (see below)
4. Activate the harvest
    a. Click 'Definitions'->'Selective Harvests' in the left menu
    b. Click 'Activate' in column 5 on the line with the <eh. name>
5. Check harvest status of the event harvest using menu "All Jobs"
    a. Click 'Harvest status'->'All Jobs' in the left menu
    b. Select "All" in "Only display job status" to the right from the menu
    c. Click the "Show" button, until the <eh. name> appears in a new job line (approx. after a minute)
    d. Check that two jobs appears and that they both have Harvest name <eh. name>
    e. Check the menu "Running jobs", that the jobs appears and that you can go to the Heritrix GUI. by clicking on the host link and by using the login/password: "admin"/"adminPassword" and close the window again.

Seed list 1 (Harvest template "default_orderxml", maxhops=0, extract_javascript=true, robots.txt=ignore, max objects=500; max bytes=400.000.000):

```
http://netarkivet.dk/adgang/
http://netarkivet.dk/in-english/
http://www.raeder.dk/
# Fjern denne linie og linien nedenunder
#http://kb-prod-udv-001.kb.dk/netarchivesuite/clock.php (is not visible from any of the harvesters in the test-system, therefore replaced for now by the link below)
http://localtimes.info/Europe/Denmark/Copenhagen/
```

Seed list 2 (Harvest template "default_orderxml", maxhops=2, extract_javascript=true, robots.txt=ignore, max objects=300; max bytes=500.000.000):

```
http://netarkivet.dk/in-english/
http://www.kaarefc.dk/
http://www.kaarefc.dk/private/
http://www.kaarefc.dk/wop/
```

Seed list "<eh. name>_default_orderxml_400000000Bytes_500Objects" for domain =raeder.dk= __"

```
http://www.raeder.dk/
```

Seed list "<eh. name>_default_orderxml_400000000Bytes_500Objects" for domain localtimes.info

```
http://localtimes.info/Europe/Denmark/Copenhagen/
```

Seed list "<eh. name>_default_orderxml_400000000Bytes_500Objects" for domain =netarkivet.dk=

```
http://netarkivet.dk/in-english/
http://netarkivet.dk/adgang/
```

Seed list "<eh. name>_default_orderxml_500000000Bytes_300Objects" for domain =netarkivet.dk=

```
http://netarkivet.dk/in-english/
```

Seed list "<eh. name>_default_orderxml_500000000Bytes_300Objects"" for domain =kaarefc.dk=

```
http://www.kaarefc.dk/
http://www.kaarefc.dk/wop/
http://www.kaarefc.dk/private/
```

4. Verify that the harvest is activated and done

This page describes how to verify that a harvest is carried out correctly

1. Click 'Harvest status'->'All Jobs' in the left menu
2. Select "All" in "Only display job status" to the right from the menu
3. Click the "Show" button, until the jobs have stepped through statuses "NEW", "SUBMITTED", "STARTED", "DONE"
4. Wait until all jobs have got status "DONE"
5. Check that you can search on Harvest name, start and end date
6. Check that you can change number of rows to be displayed per page e.g. 1 and
7. Check that you can press next and previous page and
8. Check that the reset button resets all changes to default(note that the display value is also blanked, but is 100 by default)
9. Check the following for the domain **raeder.dk**: (For example by loading the domain in the Find Domains page and clicking on "History")
    a. Check that the domain has been harvested by one job of the name <eh. name>e
    b. Check that this job has configuration **EH_default_orderxml_40000000Bytes_500Objects**
    c. Check that there is a number for 'Run number' and 'Job ID'
    d. Check that the 'Start time' and 'End time' columns approximately corresponds to time of test with the **EH** harvest
    e. Check that the 'Bytes Harvested' and 'Documents Harvested' columns contains positive numbers
    f. Check that the 'Stopped due to' columns contains "Domain-config object limit reached"
10. Check the following job details for the domain **netarkivet.dk**: (Using page SelectiveHarvests->History->Run Number 0 ->JobID 1)
    a. Check that the 'Submit time', 'Start time' and 'End time' columns approximately corresponds to time of test with **EH** harvest
    b. Click on "Browse reports for jobs"
    c. Check that you don't get any errors when you click on some of the links
    d. Click on "Browse harvest files for job"
    e. Check that you don't get any errors when you click on some of the links
    f. Click on "Browse only relevant crawl-log lines for domain netarkivet.dk"
    g. Check that you don't get any errors when you click on some of the links
11. Check the following for the domain **netarkivet.dk**: (Using page Harvest Status -> All jobs per domain)
    a. Check that the domain has been harvested by 2 jobs of the name **EH**
    b. Check that one of the jobs has configuration **EH_default_orderxml_400000000Bytes_500Objects**
    c. Check that the 'Start time' and 'End time' columns approximately corresponds to time of test with **EH**
    d. Check that one of the jobs has configuration **EH_default_orderxml_500000000Bytes_300Objects**
    e. Check that the 'Start time' and 'End time' approximately corresponds to time of test with **EH** harvest

    f. Check that 'Run number' and 'Job ID' columns contains positive numbers
    g. Check that the 'Bytes Harvested' and 'Documents Harvested' columns contains positive numbers
    h. Check that the 'Stopped due to' columns contains "Domain Completed" or "Domain-config object limit reached"

12. Check the following for the domain **kaarefc.dk**: (Using page Harvest Status -> All jobs per domain)
    a. Check that the domain has been harvested by 1 job of the name **EH**
    b. Check that the job has configuration **EH_default_orderxml_500000000Bytes_300Objects**
    c. Check that the 'Start time' and 'End time' approximately corresponds to time of test with **EH** harvest
    d. Check that 'Run number' and 'Job ID' columns contains positive numbers
    e. Check that the 'Bytes Harvested' and 'Documents Harvested' columns contains positive numbers
    f. Check that the 'Stopped due to' columns contains "Domain Completed"

5. Follow the schedule of the next job

1. Click 'Definitions'->'Selective Harvests' in the left menu
2. Check that the selective harvest <sh. name> is schedule to start in a week
3. Check that the event harvest <eh. name> is schedule to start in approx. an hour
4. Click on edit on the event harvest and override the next run time with current date and time + 5 min
5. Click 'save'
6. Click 'Harvest status'->'All Jobs' in the left menu after 5 min
7. Check that two new event harvest <eh. name> job has been generated
8. Check that NO new selective harvest <sh. name> job has been generated
9. Deactivate the event harvest <eh.name> after number two run.

6. Verify that the harvest is activated and done

1. Click 'Harvest status'->'All Jobs' in the left menu
2. Select "All" in "Only display job status" to the right from the menu
3. Click the "Show" button, until the jobs have stepped through statuses "NEW", "SUBMITTED", "STARTED", "DONE"
4. Wait until all jobs have got status "DONE"

☐ Surely the following steps are a bit superfluous? (maybe)

1. Check the following for the domain '"raeder.dk"': (Using page Harvest Status -> All jobs per domain)
    a. Check that the domain has been harvested by one job of the name <eh. name>
    b. Check that this job has configuration <eh. name>**_default_orderxml_400000000Bytes_500Objects**
    c. Check that there is a number for 'Run number' and 'Job ID'
    d. Check that the 'Start time' and 'End time' columns approximately corresponds to time of test with <eh. name> harvest
    e. Check that the 'Bytes Harvested' and 'Documents Harvested' columns contains positive numbers
    f. Check that the 'Stopped due to' columns contain "Domain Completed"
2. Check the following job details for the domain '"netarkivet.dk"': (Using page SelectiveHarvests->History->Run Number 0 ->JobID 1)
    a. Check that the 'Submit time', 'Start time' and 'End time' columns approximately corresponds to time of test with <eh. name> harvest
    b. Click on "Browse reports for jobs"
    c. Check that you don't get any errors when you click on some of the links
    d. Click on "Browse harvest files for job"
    e. Check that you don't get any errors when you click on some of the links
    f. Click on "Browse only relevant crawl-log lines for domain netarkivet.dk"
    g. Check that you don't get any errors when you click on some of the links
3. Check the following for the domain '"netarkivet.dk"': (Using page Harvest Status -> All jobs per domain)
    a. Check that the domain has been harvested by 2 jobs of the name <eh. name>
    b. Check that one of the jobs has configuration <eh. name>**_default_orderxml_400000000Bytes_500Objects**
    c. Check that the 'Start time' and 'End time' columns approximately corresponds to time of test with <eh. name>
    d. Check that one of the jobs has configuration <eh. name>**_default_orderxml_500000000Bytes_300Objects**
    e. Check that the 'Start time' and 'End time' approximately corresponds to time of test with <eh. name> harvest
    f. Check that 'Run number' and 'Job ID' columns contains positive numbers
    g. Check that the 'Bytes Harvested' and 'Documents Harvested' columns contains positive numbers
    h. Check that the 'Stopped due to' columns contains "Domain Completed"
4. Check the following for the domain '"kaarefc.dk"': (Using page Harvest Status -> All jobs per domain)
    a. Check that the domain has been harvested by 1 job of the name <eh. name>
    b. Check that the job has configuration <eh. name>**_default_orderxml_500000000Bytes_300Objects**
    c. Check that the 'Start time' and 'End time' approximately corresponds to time of test with <eh. name> harvest
    d. Check that 'Run number' and 'Job ID' columns contains positive numbers
    e. Check that the 'Bytes Harvested' and 'Documents Harvested' columns contains positive numbers
    f. Check that the 'Stopped due to' columns contains "Domain Completed"

7. Browse in data from the first event harvest only

These step require that you have your browser set up to use viewerproxy. For example in the DK test environment use the instructions at Setup DK test environment#ViewerproxySetup, or for a standalon installation use the instructions here.

1. Click 'Definitions'->'Selective Harvests' in the left menu
2. Click 'History' in column 8 on the line with the event harvest <eh. name>
3. Click 'Show jobs' in column 'Total number of jobs' on the line with 'Run number' 1

4. Click 'Select these jobs for QA with viewerproxy' (it may take some time to create page)
5. Check following in the 'Current Viewerproxy status'
6. No errors are reported
7. Check the "Currently does _not_ collect missing URLs." appear
8. Check the "Current list of missing URLs contains 0 URLs."
9. Check there is a line expressing index used from harvest <eh. name>, run 1 and built on jobs being looked at.
10. Open a New tab or window in the browser (optionally, and in same kind of browser)
11. Go to page http://netarkivet.dk/adgang/
12. Check that an error occurs saying that http://netarkivet.dk/adgang/ was not found. If this works then mark **NAS-2076** - Getting issue details... STATUS as fixed.
13. Go to page http://www.kaarefc.dk
14. Check that this page contains data
15. Go to page http://www.kaarefc.dk/wop/
16. This page should exist.
17. Go to page http://indvandrerbiblioteket.dk
18. Check that an error occurs saying that www.indvandrerbiblioteket.dk was not found
19. Go to page `http://localtimes.info/Europe/Denmark/Copenhagen/`
20. Check that a page containing date and time of the first harvest appears

8. Browse in data from the second event harvest only

1. Click 'Definitions'->'Selective Harvests' in the left menu
2. Click 'History' in column 8 on the line with event harvest <eh. name>
3. Click 'Show jobs' in column 'Total number of jobs' on the line with 'Run number' 2
4. Click 'Select these jobs for QA with viewerproxy' (it may take some time to create page)
5. Check following in the 'Current Viewerproxy status'
6. No errors are reported
7. Check the "Currently does _not_ collect missing URLs." appear
8. Check the "Current list of missing URLs contains 0 URLs."
9. Check there is a line expressing index used from harvest <eh. name>, run 2 and built on jobs being looked at.
10. Open a New tab or window in the browser (optionally, and in same kind of browser)
11. Go to page http://www.netarkivet.dk
12. Check that an error occurs saying that www.netarkivet.dk was not found. If this works then mark **NAS-2076** - Getting issue details... STATUS as fixed.
13. Go to page http://www.kaarefc.dk
14. Check that this page contains data
15. Click on a local link (e.g. =http://www.kaarefc.dk/wop/ in link for= 'Here').
16. Check that this page contains data
17. Go to page http://indvandrerbiblioteket.dk
18. Check that an error occurs saying that www.indvandrerbiblioteket.dk was not found
19. Go to page `http://localtimes.info/Europe/Denmark/Copenhagen/`
20. Check that a page containing date and time of the second harvest appears (Note: "Refresh" may be necessary)

9. Browse in data from the selective harvest only

1. Click 'Definitions'->'Selective Harvests' in the left menu
2. Click 'History' in column 8 on the line with the selective harvest &lt;sh. name>
3. Click 'Show jobs' in column 'Total number of jobs' on the line with 'Run number' 0
4. Click 'Select these jobs for QA with viewerproxy' (it may take some time to create page)
5. Check following in the 'Current Viewerproxy status'
    a. No errors are reported
    b. Check the 'Currently does _not_ collect missing URLs.' appear
    c. Check the 'Current list of missing URLs contains 0 URLs.'
    d. Check there is a line concerning index used for harvest &lt;sh. name>, run 0 and built on jobs being looked at.
6. Open a new tab or window in the browser (optionally, and in same kind of browser)
7. Go to page http://mazda.dk
8. Check that this page contains data and all links are functional
9. Go to a random internet page not on http://netarkivet.dk (but not https). The page should NOT be found. (Example: http://www.pligtaflevering.dk)

10. Verify that data is deduplicated

1. Click on the JobID for your second finished event harvest <eh-name> in the Job status overview
2. Click on "Browse reports for jobs"
3. Click on the "processors-report" e.g. "metadata://netarkivet.dk/crawl/reports/processors-report.txt?heritrixVersion=3.3.0-LBS-2014-03 &harvestid=1&jobid=1" (or similar. The harvestid and jobid will probably differ)
4. Check that there is a deduplicator processors-report similar to this one (the numbers will be different), but duplicates found should be non-zero:

```
Total handled: 88
Duplicates found: 20 20.0%
Bytes total: 6391852 (6.1 MB)
Bytes discarded: 0 (0 0.0%
New (no hits): 88
Exact hits: 0
Equivalent hits: 0
.....
```

5. Check also the deduplicator report for the first run of the event harvest. The number of duplicates should be zero.

11. Define and run low bandwidth selective harvest

(The idea behind this is to create a job that is slow enough that one has time to terminate it before it is finished.)

1. Go to Edit Harvest Templates page. Download default_orderxml.
2. Edit it to replace disposition.maxPerHostBandwidthUsageKbSec with 30.
3. Upload it as a new config: default_orderxml_low_bandwidth
4. Go to edit-page for domain 'netarkivet.dk', edit defaultconfig, and replace harvesttemplate with 'default_orderxml_low_bandwidth'
5. Make a new selective harvest definition with a name you can remember
    a. Click 'Definitions'->'Selective Harvests' in the left menu
    b. Click 'Create new harvest definition' in the bottom of the main window
    c. Fill in the Harvest name and note the name for later use (from now referred as <sh1. name>)
    d. Choose "Once_a_week" in the drop down list for 'Schedule'
    e. Write =netarkivet.dk= in the 'Enter Domain...' window and click 'Add domains'
    f. Click 'Save'
6. Activate the selective harvest
    a. Click 'Activate' in column 5 on the line with the <sh1. name>
    b. Check that the time in the "Next Run" column time on the line with the <sh1. name> is now.
7. Check harvest status of the selective harvest
    a. Click 'Harvest status'->'All Jobs' in the left menu
    b. Select "All" in "Only display job status" to the right from the menu
    c. Click the "Show" button, until the <sh1. name> appears in a new job line (approx. after a minute)
    d. Check that the job has status "NEW", it may have turned into status "SUBMITTED" or status "STARTED" before you see it.
8. Check job creation in the system status for the selective harvest
    a. Click 'Systemstate'->'Overview of the system state'
    b. Find and click 'HarvestJobManagerApplication' in the 'Application' column for the KB kb-test-adm-001
    c. Click 'show all' in the "Index" header
    d. Check that there exists a line with the message "INFO: Created 1 jobs for harvest definition and a line after that "INFO: Job #1 submitted, and later the line: "INFO: Job #1 has been started by the harvester."

12. Terminate a running harvest

Use the H3 Remote Access section under Harvest status to terminate the job once it has started harvesting.

13. Check the Heritrix terminated job is logged in the Job details in ADM GUI

1. Click on refresh until the job disappears in 'System Overview' (5 min.) (ie you see "Starts to listen to new jobs" on the HarvestControllerApplication where the job was running)
2. Click 'Harvest status' and select your terminated job by clicking on the Job ID number
3. Verify that under the 'Included domains and configurations' some domains are "Stopped due to": "Harvesting aborted" (Should be only domain: netarkivet.dk)

14. Start a snapshot harvest with max 1.000.000 bytes

1. Make a new snapshot harvest definition with a name you can remember
    a. Click 'Definitions'->'Snapshot Harvests' in the left menu
    b. Click 'Create new harvestdefinition' in the bottom of the main window
    c. Fill in the 'Harvest name' and note the name for later use (from now referred as <snh. name>)
    d. Set `Max number of bytes per domain` to 1000000 (1 Mbytes)
    e. Click Save
    f. Click 'Activate' in column 4 on the line with the <snh. name>
2. Check scheduling of jobs
    a. Click 'Harvest status'->'All Jobs' in the left menu
    b. Select to view NEW jobs

c. Check that a new snapshot harvest <snh. name> job has been generated (may take a minute before jobs appear)
d. Click 'Systemstate' in the left menu
e. Check that the HarvestJobManager application contains the message "INFO: Created X jobs for harvest definition" (choose Application HarvestJobManager and Show all lines)
f. Check That there are no warnings on the different applications

15. Terminate a Job

Terminate the job harvesting netarkivet.dk in the snapshot harvest. Wait for the other jobs to finish.

17. Start a snapshot harvest with max 100000 bytes

1. Make a new snapshot harvest definition with a name you can remember
    a. Click 'Definitions'->'Snapshot Harvests' in the left menu
    b. Click 'Create new harvestdefinition' in the bottom of the main window
    c. Fill in the 'Harvest name' and note the name for later use (from now referred as <snh. name.2>)
    d. Set 'Max number of bytes per domain' to 100000.
    e. Click on the <snh.name> under 'Harvest only domains that were not completely harvest in a previous harvest'
    f. Click Save
    g. Click 'Activate' in column 4 on the line with the <snh. name.2>
2. Check scheduling of jobs
    a. Click 'Harvest status'->'All Jobs' in the left menu
    b. Select to view NEW jobs
    c. Check that a new snapshot harvest <snh. name.2> job has been generated (may take a minute before jobs appear)
    d. Click 'System status' in the left menu
3. Verify job status
    a. Click 'Harvest status'->'All Jobs' in the left menu
    b. Select "All" in "Only display job status" to the right from the menu
    c. Click the "Show" button, until the jobs have stepped through statuses "NEW", "SUBMITTED", "STARTED", "DONE"

18. Check that the domains stopped by Heritrix termination are not part of the next harvest.

1. Click on refresh until the job disappears in system overview (about 5 min.)
2. Click 'Harvest status' and for each jobs in the snapshot harvest
    a. Click on Job ID number
    b. Verify that the 'Included domains and configurations' are without the domains which was stopped due to 'Harvesting aborted' in the previous harvest, and the rest are 'Domain Completed' or 'Max Bytes limit reached'