User Manual

Printer friendly version

Introduction

The NetarchiveSuite is a software toolset built to be operated by librarians with a minimum of training in basic internet concepts, webcrawling methods and usage of this system. It's built to handle the national obligations of preserving the internet in Denmark and its design and architecture is much influenced by the Legal Deposit Law stating the three different harvest types:

- snap shot harvests of the national domain
- selective harvests of selected websites
- event harvests in connection with national events

It's also designed to be distributed across the world due to the fact that the danish installation is driven by the two national libraries: The State and University Library, Aarhus and the Royal Library in Copenhagen.

The basic concept in the NetarchiveSuite harvesting module is the notion of domains.

A domain has a two part name (e.g. netarchive.dk) or is an IP-number. For countries with other domain-name structures (e.g. UK), the system will need minor changes to apply to those naming conventions.

A domain can hold multiple so called configurations. A configuration describes how to harvest the domain or a part of the domain. So one configuration could harvest the whole domain (used by the snapshot functionality) and other configurations could take different minor parts of the same domain (for the selective / event harvest).

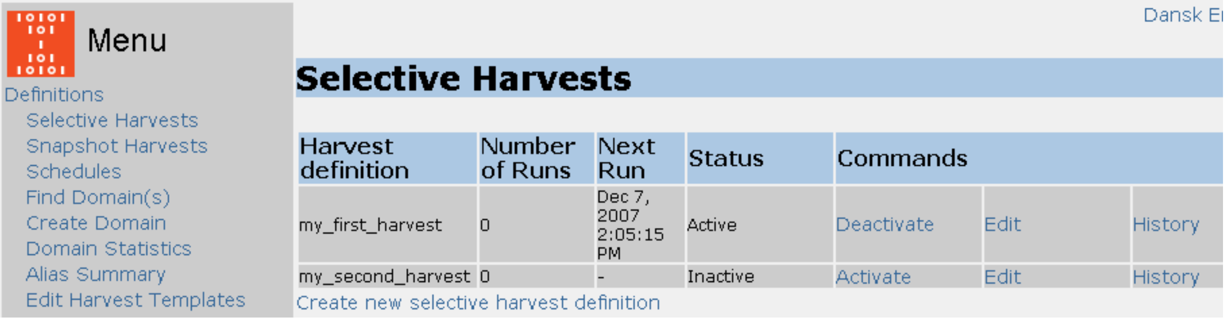
A configuration consists basically of two things

- A harvester template (predefined templates for the Heritrix web crawler)
- A number of seedlists to use with that template

A domain will always have a default configuration (selectable) and that configuration will be used when starting a snap shot harvest. The snap shot harvest therefore takes all domains known in the database.

edit

Selective Harvests



Selective Harvests

Harvest definition	Number of Runs	Next Run	Status	Commands		
my_first_harvest	0	Dec 7, 2007 2:05:15 PM	Active	Deactivate	Edit	History
my_second_harvest	0	-	Inactive	Activate	Edit	History

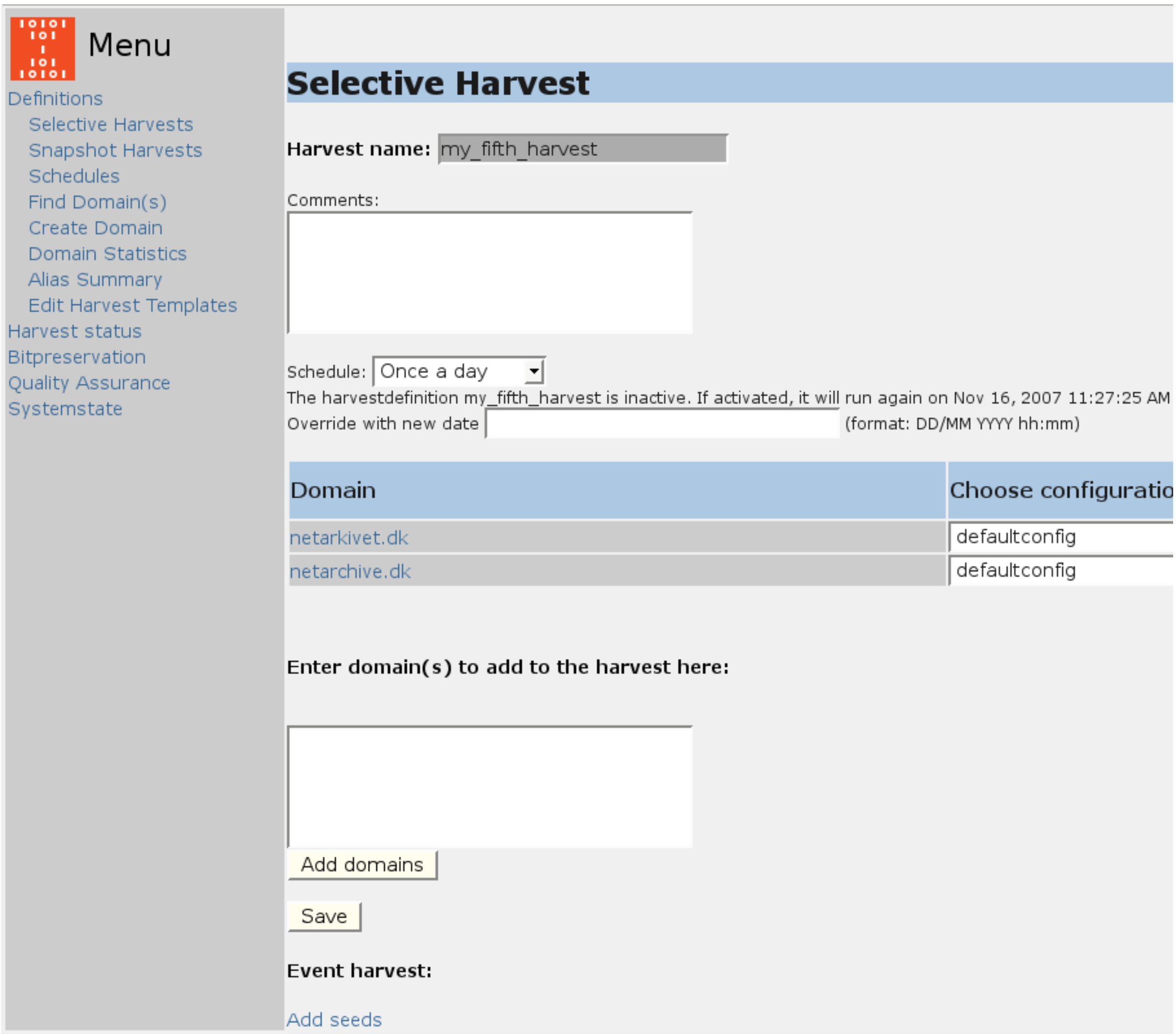
The front page by default shows the list of selective harvests.

You can [***Activate***] *an inactive harvest definition and* [**Deactivate**] *an active harvest definition. If you deactivate a running harvest, the system will finish the running jobs.*

Click on [***Edit***] *to change an existing harvest definition or* [**Create a new harvest definition**]

Click on [***History***] if you wish to trace back all the jobs from former finished harvests.

Creating/editing a selective harvest



Selective Harvest

Harvest name: my_fifth_harvest

Comments:

Schedule: Once a day

The harvestdefinition my_fifth_harvest is inactive. If activated, it will run again on Nov 16, 2007 11:27:25 AM

Override with new date (format: DD/MM YYYY hh:mm)

Domain	Choose configuratio
netarkivet.dk	defaultconfig
netarchive.dk	defaultconfig

Enter domain(s) to add to the harvest here:

Event harvest:

Create a new selective harvest definition by pressing [***Create new selective harvestdefinition***] from the frontpage.

Give the harvestdefinition a recognizable harvest name – you can not change it later. If necessary add a comment.

Choose a schedule from the dropdown list.

Now you can add domains to the harvestdefinition.

Write the name of the domains you want to add in the box "**Enter domain(s) to add to the harvest here**" and click on [***Add domains***].

*The added domains will appear in the column "**Domain**".*

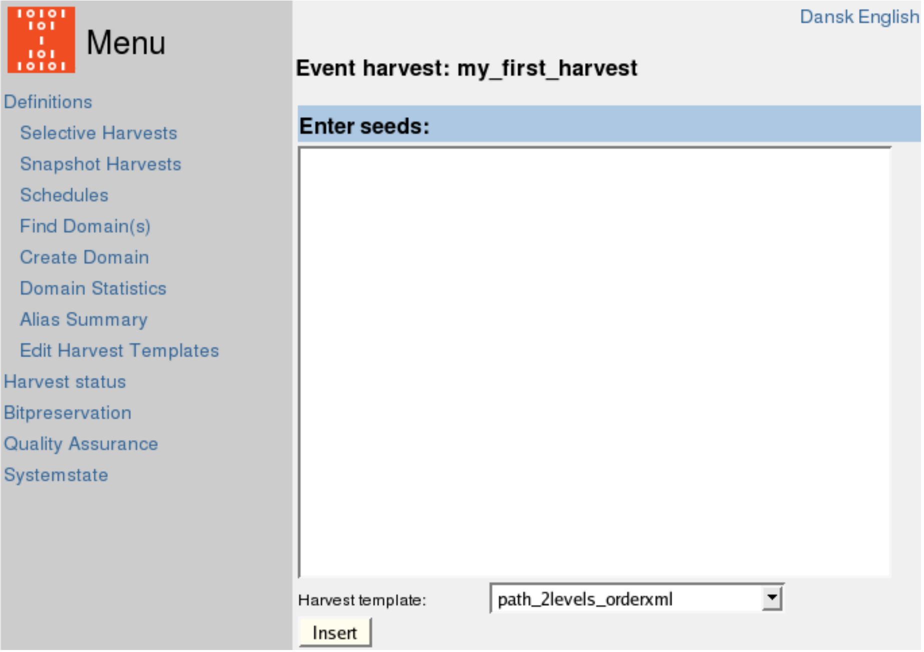
For each added domain, choose the wanted configuration from the dropdown list for each domain. Press ***[Save]*** to save the harvestdefinition.

The scheduling of selective harvest definitions can be overridden by filling out the input field **Override with new date**. Simply set the date to whenever you wish the harvest definition to run next time. The scheduling of the harvest definition will continue from that point in time.

Event harvest

Event harvests are treated almost the same as selective harvests in the system. The only difference is a power-adding of domains function. This could be used for selective harvests as well but was developed for event harvesting definitions where the operator must fill in larger number of URLs without having to edit configurations and seedlists on all those domains.

Adding seeds to an event harvest



Event harvest: my_first_harvest

Enter seeds:

Harvest template: path_2levels_orderxml

Insert

Use [***Add seeds***]. Enter identified start-URLs covering the event in the "Enter seeds:" box. Select a harvest template with the "Harvest template" drop down box.

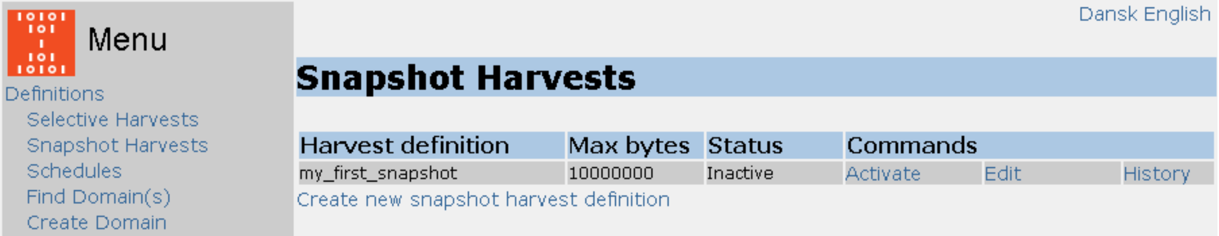
All seeds will use the same template, so to harvest different seeds with different templates you need to add them bunch by bunch for each template you need for your event harvest.

Pressing [***Insert***] starts the power-adding function. This function runs through the entered seeds one by one and does the following with each seed:

1. Finds the domain from which the seed derives
2. Creates a seedlist with the name of the harvestdefinition and the template as seedlist-name
3. Creates a configuration with the name of the harvestdefinition and the template as configuration-name. And select the seedlist from (2) to use with the new configuration.If the seedlist to create in (2) or the configuration to create in (3) already exist (If the power-adding function has been used before with other seeds from the same domain in the same event harvest) the system will only add the new URLs to the existing seedlist.

edit

Snapshot Harvests



Snapshot Harvests

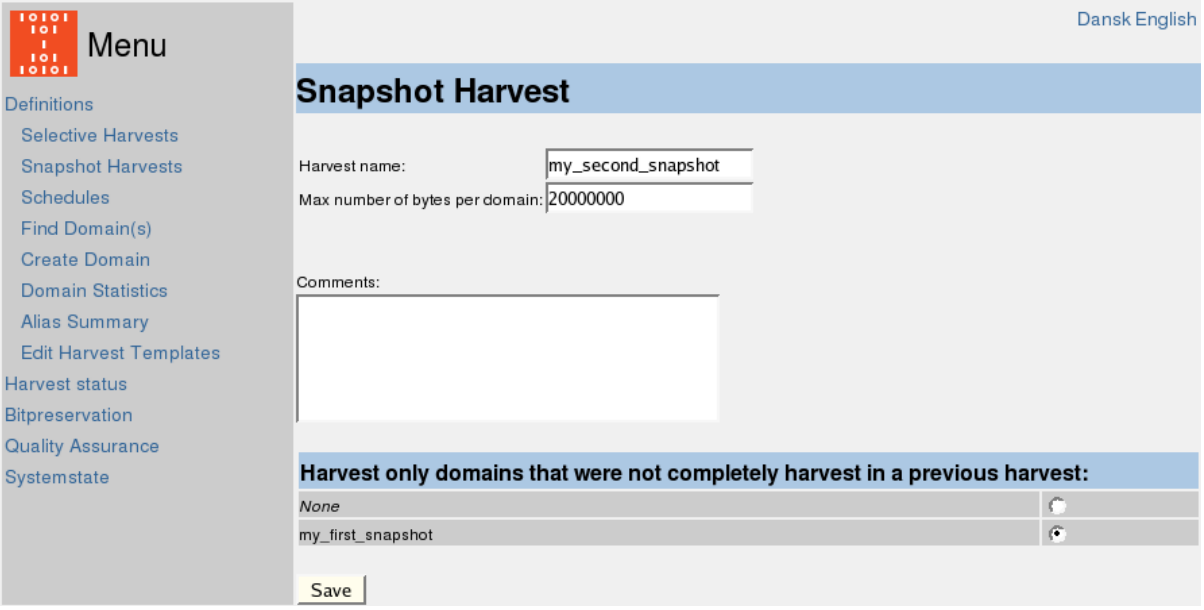
Harvest definition	Max bytes	Status	Commands		
my_first_snapshot	10000000	Inactive	Activate	Edit	History

Create new snapshot harvest definition

On [***Snapshot Harvests***] new snapshot harvests are started, harvesting all domains known to the system in their default configurations. An overview of all snapshot harvests is also provided.

[***Create new harvestdefinition***] opens the template below.

Creating/editing a snapshot harvest



Snapshot Harvest

Harvest name: my_second_snapshot

Max number of bytes per domain: 20000000

Comments:

Harvest only domains that were not completely harvest in a previous harvest:

None	○
my_first_snapshot	●

This page is used to define name and size (max. bytes per domain) of the harvest. It is recommended to systematize the naming for clarity, e.g. 2007-1, 2007-2 etc.

The size of the harvest can be defined in two ways: at the harvest definition [***Snapshot Harvests***] or at the configuration of the single domain. It will always be the lower size limit stopping the harvesting of a domain.

Comments can freely be added.

Snapshot harvests can be based on previous snapshots in the sense that it can be limited to only harvest domains that hit the max number of bytes limit in a previous harvest.

The domains completely finished (not hitting the max number of bytes limit – either on the configuration level or on the snapshot harvest level) in the first harvest will not be included in the second.

All other domains will be harvested from the beginning up till the newly specified limit for the new snapshot.

[***Save***] saves the harvest definition and returns to [***Snapshot harvests***].

After defining a snapshot harvest the harvest is activated with the [***Activate***] button on the snapshot frontpage. Harvest will not start until you press [***Activate***]. Status then changes to ‘Active’.

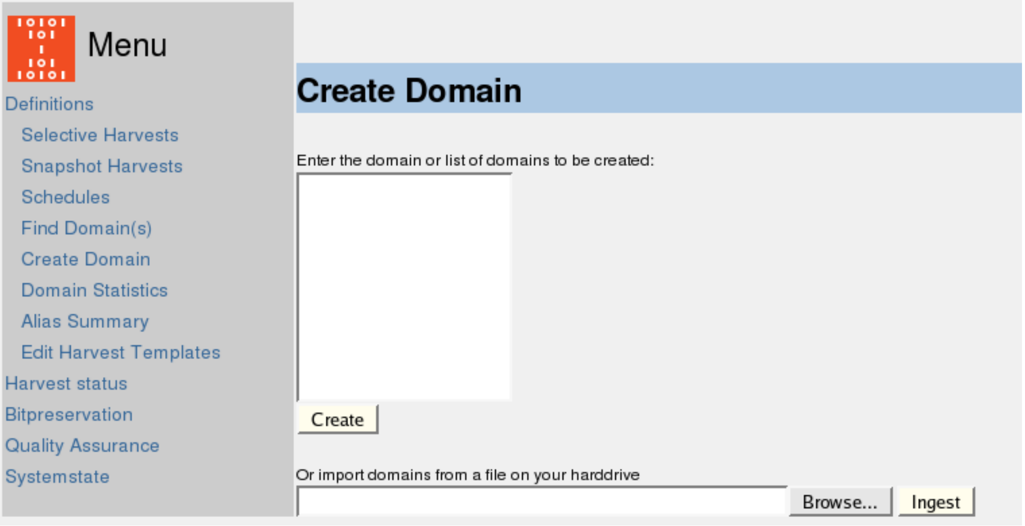
[***Deactivate***] is not relevant in Snapshot Harvests because they only run once. By [***Edit***] the Snapshot Definition can be changed but only before activation. Parameters changeable are size, commentary and if previous harvest startpoint should be used. The name can not be changed.

[***History***] provides an overview of the specific harvest: see User Manual/Harvest History

edit

Domains

Creating Domains



Create Domain

Enter the domain or list of domains to be created:

Or import domains from a file on your harddrive

The [***Create domain***] is used for creating new domains in the system. It is possible to create a single domain as well as list of domains. It is also possible to import domains from a file.

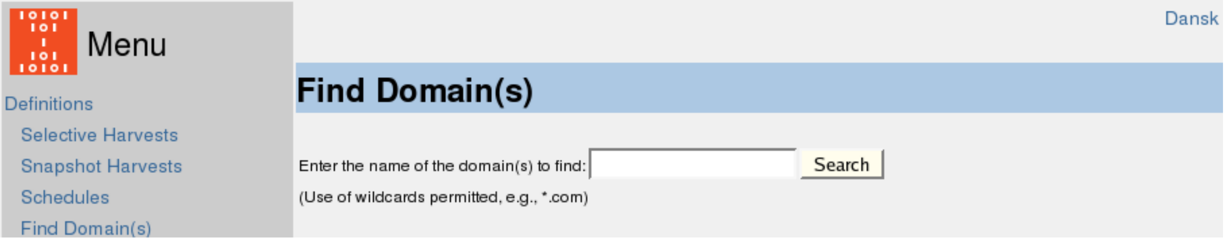
To create single domains enter domain names in the text box and press [***Create***].

To bulk create domains from a file select the file from your local computer with [***Browse***] and press [***Ingest***]. The file must be a simple list of domain names – one at each line. The file must be UTF-8 encoded if it contains special characters.

New domains get a default configuration when created (with the defaultorderxml template and a maximum number of bytes at 500Mbytes). New domains also get a defaultseedlist when created. This seedlist contains a link to the frontpage of the domain. The frontpage is defined as www.domainname – e.g. ' www.netarchive.dk' for the domain 'netarchive.dk'

Already existing domains in the system will not be recreated.

Finding Domains



Find Domain(s)

Enter the name of the domain(s) to find:

(Use of wildcards permitted, e.g., *.com)

[***Find Domain(s)***] is used to find domains existing in the system.

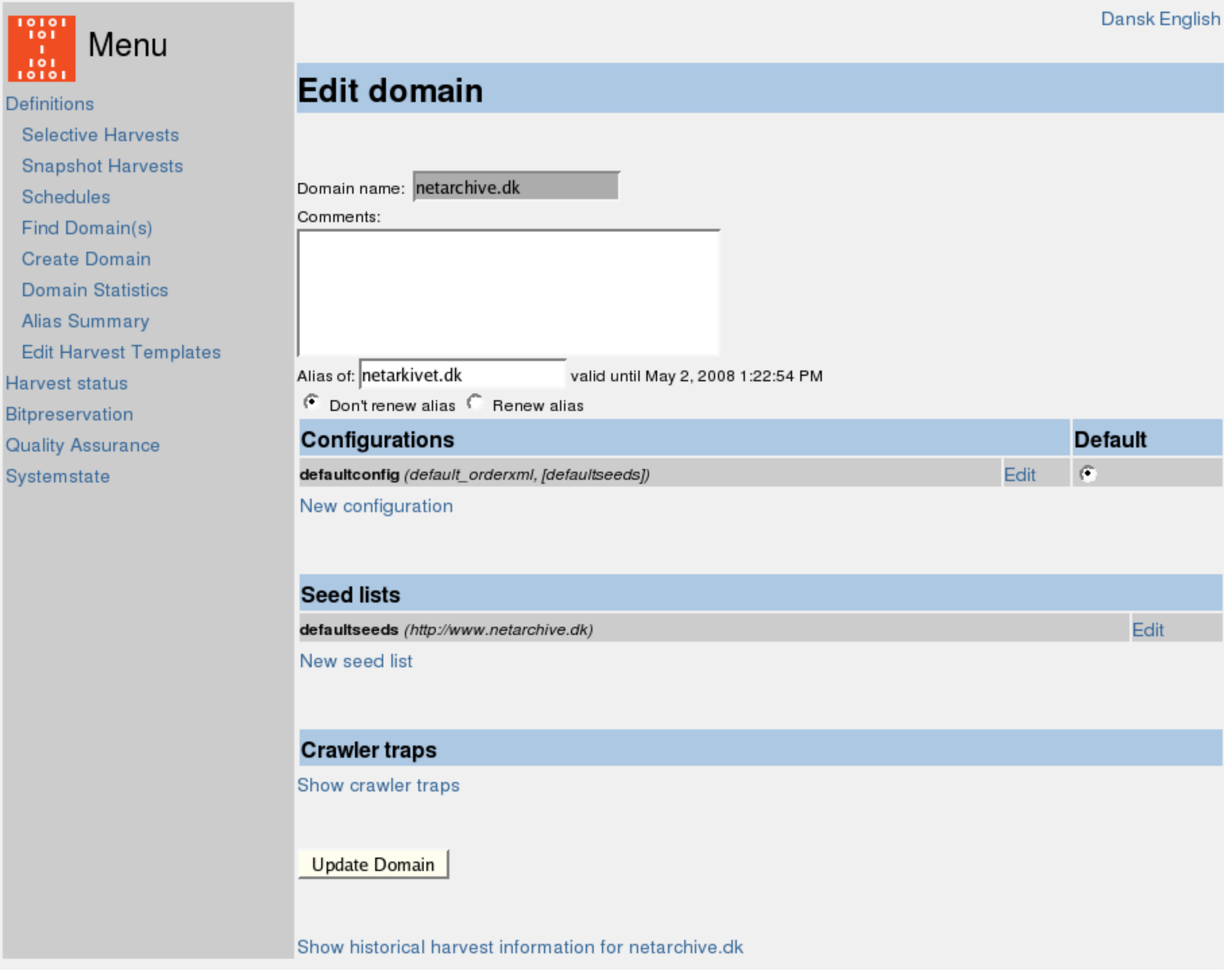
Write a domain name in the box (e.g. kb.dk). Searching is done on the complete text string. Press [***Search***].

Left and/or right wildcards with *.

If there are several hits, a list is given of the found domains. If only one hit it leads directly to the Domain page.

If the search for a specific domain results in no hits, you are prompted with the ability to create the domain in the system and by accepting [***Yes***] it leads directly to Domain page for the newly created domain.

Editing Domains



Edit domain

Domain name: netarchive.dk

Comments:

Alias of: netarkivet.dk valid until May 2, 2008 1:22:54 PM

● Don't renew alias ○ Renew alias

Configurations		Default
defaultconfig *(default_orderxml, [defaultseeds])*	Edit	●

Seed lists	
defaultseeds *(http://www.netarchive.dk)*	Edit

Crawler traps

’***Edit domain***’ is an overview of a single domain where it is possible to edit the domain’s definition in the harvest system.

Free commentary text box.

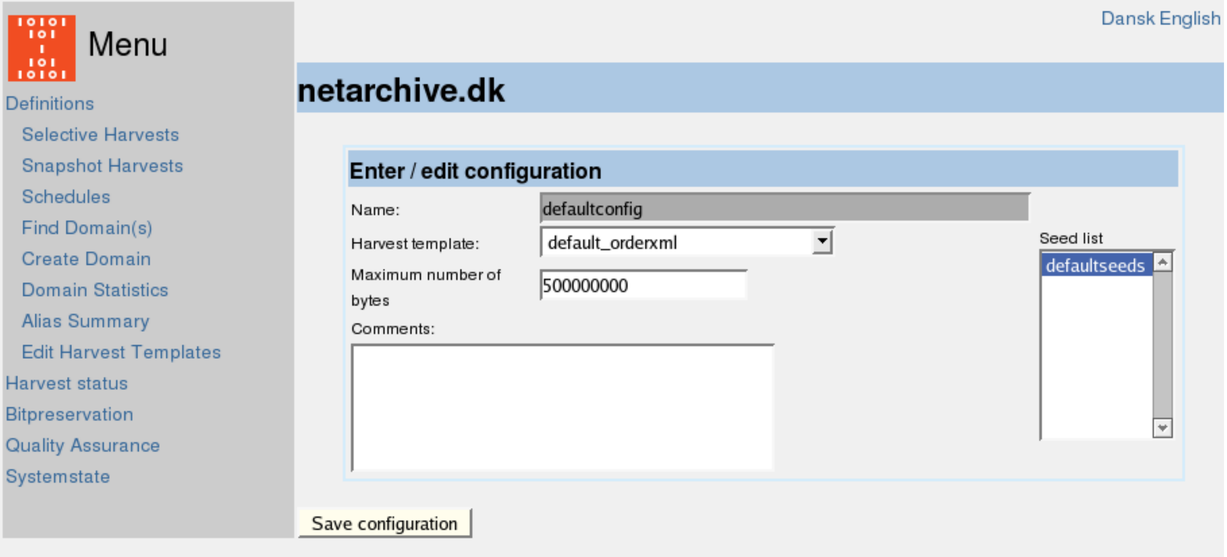
’Alias of’: Here it can be stated if the domain is an alias of another domain – they are 100% identical in content and only one of them should be harvested. Domains marked as an alias will not be harvested within the snapshot harvests. Alias is defined one year at a time and then has to be renewed.

‘Configurations’: [***New configuration***] and [***Edit***] open a new page: ‘***Enter/edit configuration***’ (see below) ‘Seed lists’: [***New seed list***] and [***Edit***] opens a new page: ‘***Enter/edit seed list***’. (see below)

‘Crawler traps’: [***Show crawler traps***] Open a new text box: ‘***Crawler traps***’ (see below)

[***Show historical harvest information for …***] *opens a new page “***Harvest history for domain....***” (see User Manual/Harvest History) ['*

Editing configurations



Enter / edit configuration

Name: defaultconfig

Harvest template: default_orderxml

Maximum number of bytes: 500000000

Comments:

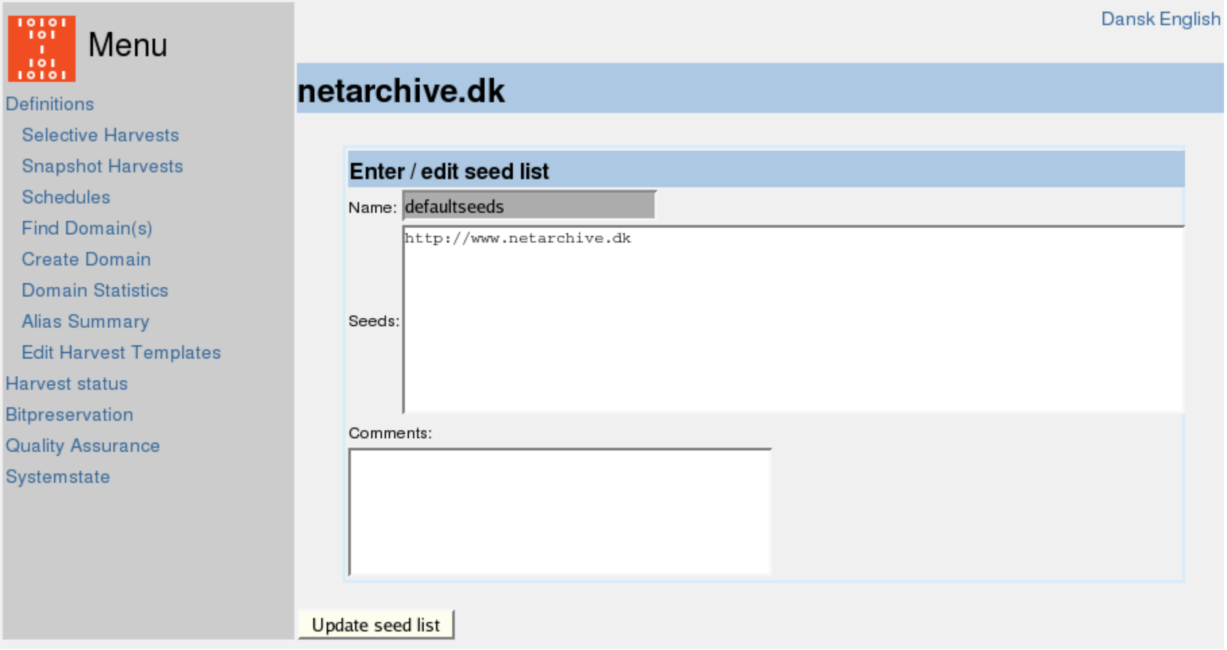
Seed list: defaultseeds

’Enter/edit configuration’ is used to define a new configuration and edit an existing one. A configuration contains information about which Harvest template and Seed lists are used (more than one Seed list can be used - hold down CTRL).

At the creation of a new configuration a name is given that thereafter can not be changed.

Furthermore it is possible to choose between different Harvester templates and maximum number of bytes to be harvested in each harvest of the configuration. The default number of bytes is 500Mbytes

Editing seed lists



Enter / edit seed list

Name: defaultseeds

Seeds:

```
http://www.netarchive.dk
```

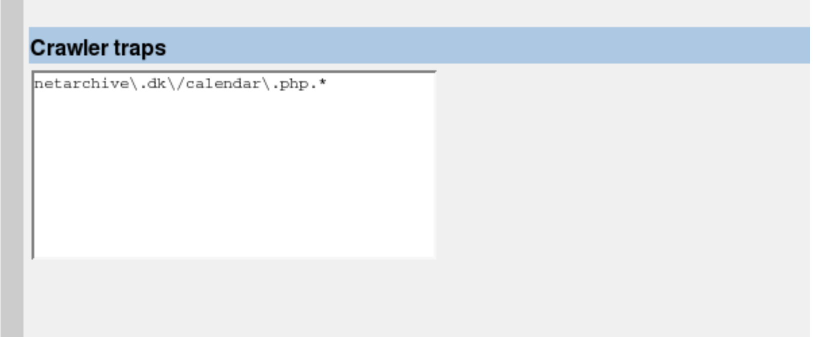
Comments:

’Enter/edit seed list’ is used to define a new Seed list or to edit an existing one.

At the creation of a new Seed list a name is given that thereafter can not be changed.

In the ’Seeds’ text box a list of seeds to be harvested is given. Seeds can be omitted by writing a # prefix, e.g. # http://www.kb.dk. This can also be used for comments inside the seedlist – e.g. '#this seed is important'

Editing crawlertraps



Crawler traps

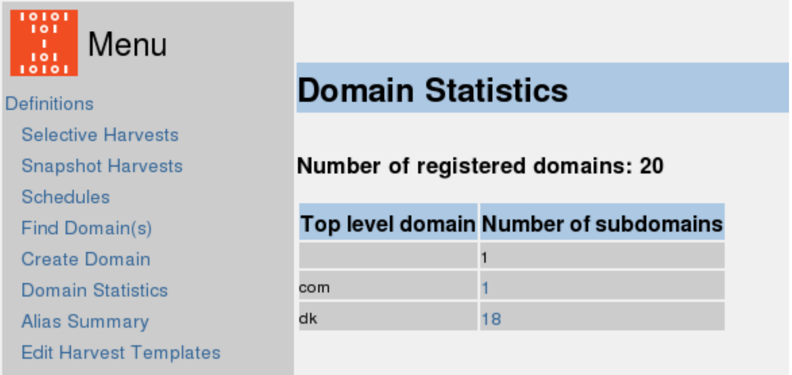
```
netarchive\.dk\/calendar\.php.*
```

A crawlertrap is a path followed blindly by the harvester which in principle can continue forever. This typically could be a calendar.

To avoid crawlertraps on a domain, the administrator can state parts of URLs that should never be harvested (in any configuration). Matching URLs are omitted in all harvests of the domain and in other domains harvested in the same job. So be careful not to give too general statements that could potentially omit things on other domains (perhaps always include the domainname itself in the statement).

The string of text must be stated as a regular expression.

Domain statistics



Domain Statistics

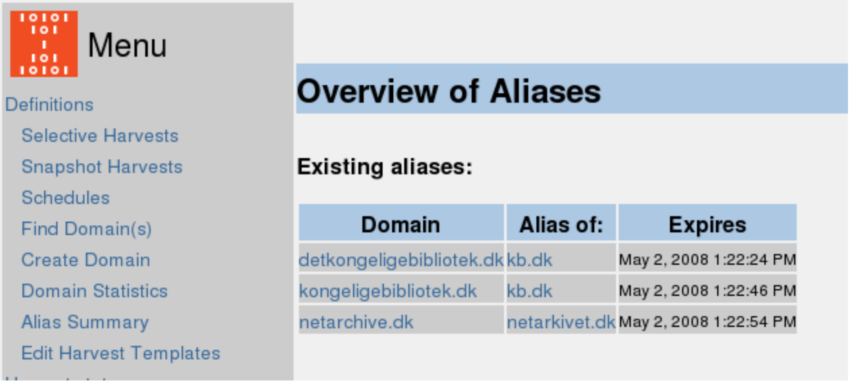
Number of registered domains: 20

Top level domain	Number of subdomains
	1
com	1
dk	18

The domain statistics page will give you information about number of subdomains for each unique Top level domain known in the system. IP-numbers will be counted separately.

The number in the “Number of subdomains” column is clickable and will do a search for all domains matching that Top level domain. This is only applicable to Top level domains with a limited number of subdomains since the matching domains will be listed on one page – and that page will get very long if the system contains hundreds of thousands of domains.

Alias summary



Overview of Aliases

Existing aliases:

Domain	Alias of:	Expires
detkongeligebibliotek.dk	kb.dk	May 2, 2008 1:22:24 PM
kongeligebibliotek.dk	kb.dk	May 2, 2008 1:22:46 PM
netarchive.dk	netarkivet.dk	May 2, 2008 1:22:54 PM

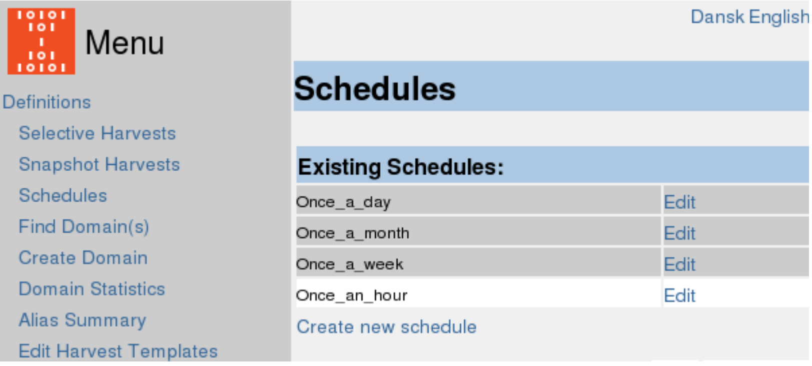
The alias summary page gives an overview of the domains marked as aliases of other domains in the system. Both domain names are clickable and will open the domain page for the clicked domain.

The “Expires” column shows when the alias expires (12 month after they are marked). The mark does not disappear after 12 month in the database but the “Overview of Aliases” page will show the “expired” ones in the top.

To renew an alias for another 12 month one is currently forced to open the domain page of the marked domain (the “Domain” column) – select “renew alias” and press “Save”

edit

Schedules



Schedules

Existing Schedules:

Once_a_day	Edit
Once_a_month	Edit
Once_a_week	Edit
Once_an_hour	Edit

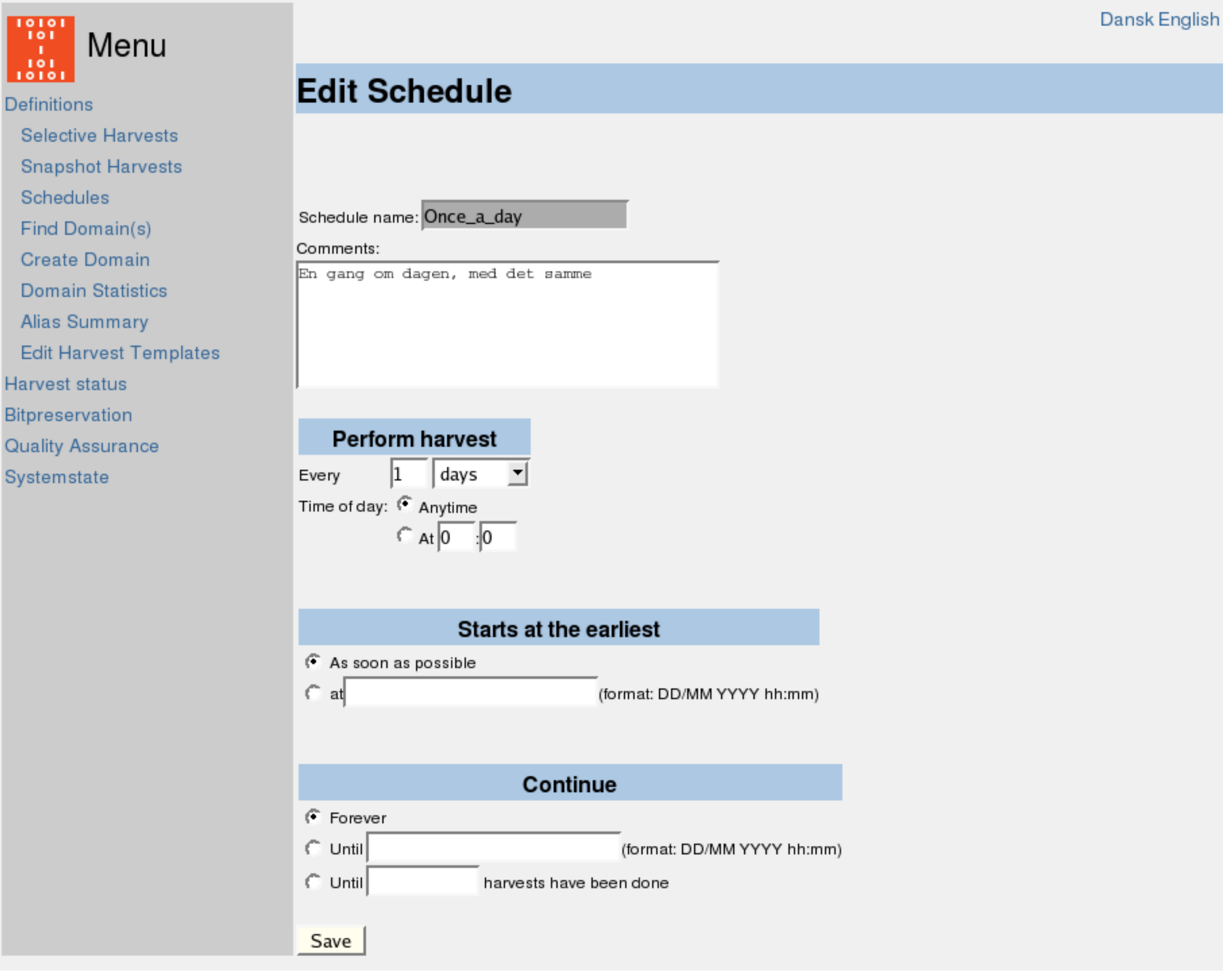
Create new schedule

Schedules are only applied on selective and event harvests.

A schedule defines a harvesting frequency. The minimum entity is one hour. It is possible to choose an automatically fixed start and/or end time for a specific harvest.

It is possible to create an infinite number of schedules.

For a new schedule click on [***Create new schedule***]*. And to edit an existing schedule press [***Edit***]*.*



Edit Schedule

Schedule name: Once_a_day

Comments: En gang om dagen, med det samme

Perform harvest

Every 1 days

Time of day: ● Anytime

○ At 0 :0

Starts at the earliest

● As soon as possible

○ at (format: DD/MM YYYY hh:mm)

Continue

● Forever

○ Until (format: DD/MM YYYY hh:mm)

○ Until harvests have been done

Save

Give the schedule an easily recognizable name – note that it can't be changed once saved. If necessary, add a comment.

Fill in the frequency and – if necessary - time of the day for the harvest to run.In the drop down menu you have the choice between hours, days, weeks and months. Changing “***days*** ” *switches the* “**Time of day**” *so that:*

- “hours” lets you choose a specific minute of the hour
- “days” lets you choose a specific time of day
- “weeks” lets you choose a specific day of the week
- “months” lets you choose a specific day of the month

After selecting the frequency you must select “***Start at the earliest***” which could either be as soon as possible (default) or at a specific date and time.

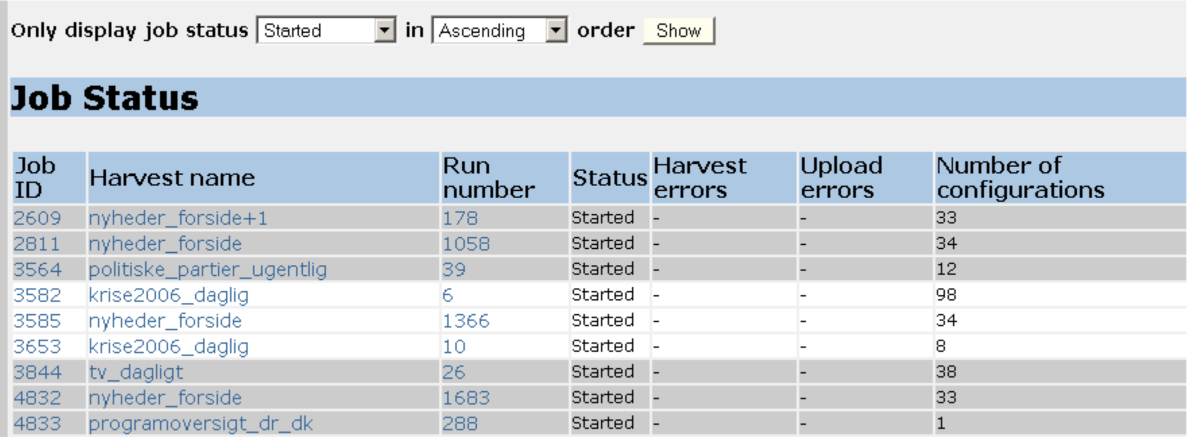
The last thing to determine is how long this schedule should go on. The default for the duration of a schedule is forever. It is also possible to choose an end date and a certain number of harvests to perform. This allows you to define schedules that will only run in a shorter period – e.g. in connection with an event harvest where the date range in which to harvest is predefined.

edit

Harvest History

All jobs

*[**Harvest Status**] in the left menu by default shows a chronological list of all jobs ever harvested with status **Started** in **ascending order**. The same does [**All jobs**]*



Only display job status Started ▾ in Ascending ▾ order Show

Job Status

Job ID	Harvest name	Run number	Status	Harvest errors	Upload errors	Number of configurations
2609	nyheder_forside+1	178	Started	-	-	33
2811	nyheder_forside	1058	Started	-	-	34
3564	politiske_partier_ugentlig	39	Started	-	-	12
3582	krise2006_daglig	6	Started	-	-	98
3585	nyheder_forside	1366	Started	-	-	34
3653	krise2006_daglig	10	Started	-	-	8
3844	tv_dagligt	26	Started	-	-	38
4832	nyheder_forside	1683	Started	-	-	33
4833	programoversigt_dr_dk	288	Started	-	-	1

If information is wanted for jobs with other statuses (or ***All*** statuses) or other sort order, then this can be specified in the combo-boxes in the top of the page and then activated by clicking the ***Show*** button.

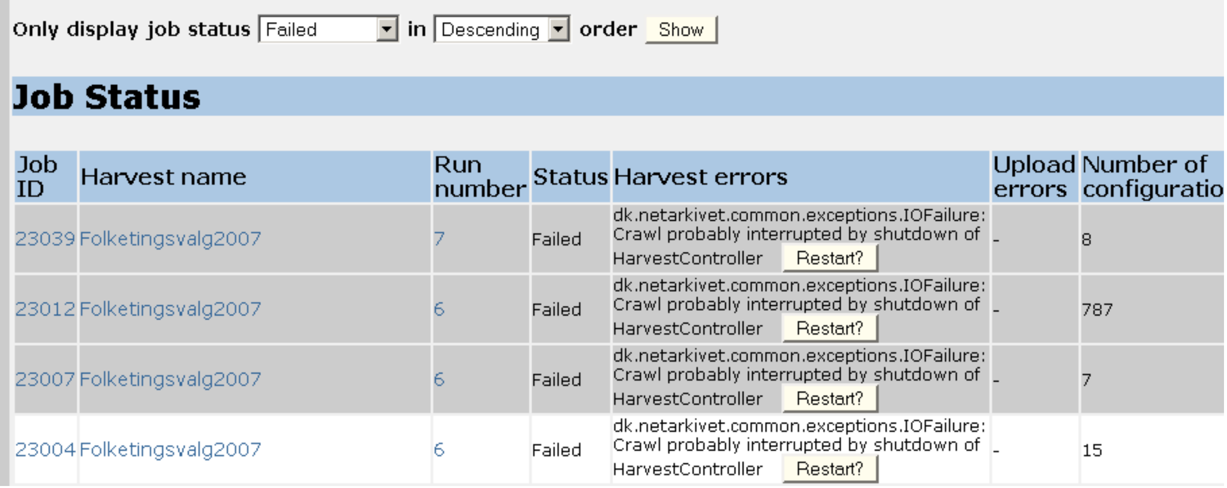
For each job the page shows information about the job and its status as well as information about errors (harvest errors or upload errors) and number of configurations in the job.

Chose ***[Run number]*** *if you want to check details on a specific run of that harvestdefinition – note that a run can consist of multiple jobs.*

Chose ***[Harvest name]*** *if you want to check details on the history of a specific harvest definition.*

Chose ***[JobID]*** *if you want to check details on a specific job.*

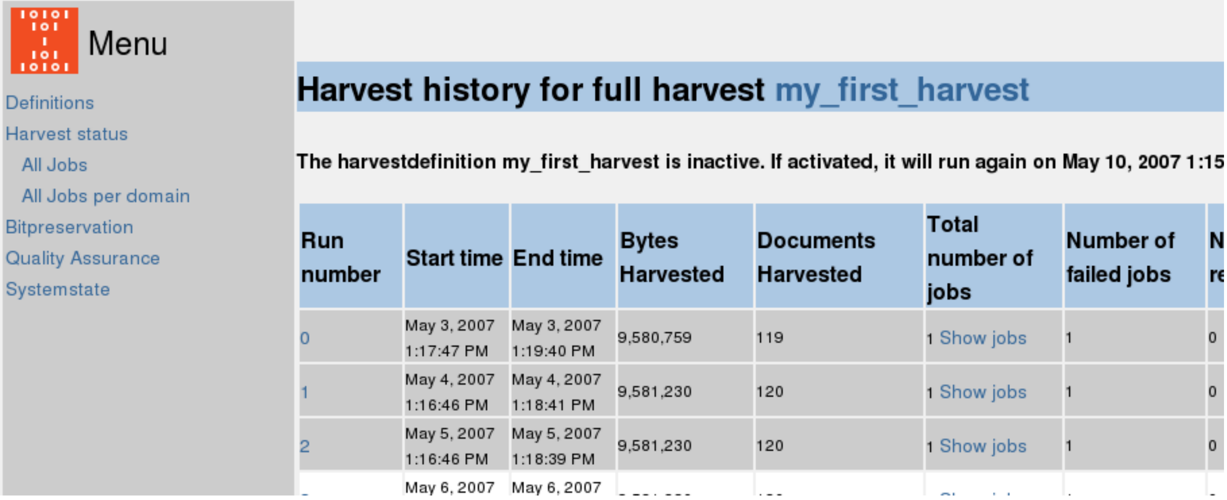
In case of Harvest errors, a [***Restart***] button will appear and the operator can choose to resubmit that specific job to be harvested again.



Job Status

Job ID	Harvest name	Run number	Status	Harvest errors	Upload errors	Number of configuratio
23039	Folketingsvalg2007	7	Failed	dk.netarkivet.common.exceptions.IOFailure: Crawl probably interrupted by shutdown of HarvestController Restart?	-	8
23012	Folketingsvalg2007	6	Failed	dk.netarkivet.common.exceptions.IOFailure: Crawl probably interrupted by shutdown of HarvestController Restart?	-	787
23007	Folketingsvalg2007	6	Failed	dk.netarkivet.common.exceptions.IOFailure: Crawl probably interrupted by shutdown of HarvestController Restart?	-	7
23004	Folketingsvalg2007	6	Failed	dk.netarkivet.common.exceptions.IOFailure: Crawl probably interrupted by shutdown of HarvestController Restart?	-	15

History of a harvestdefinition



Harvest history for full harvest my_first_harvest

The harvestdefinition my_first_harvest is inactive. If activated, it will run again on May 10, 2007 1:15

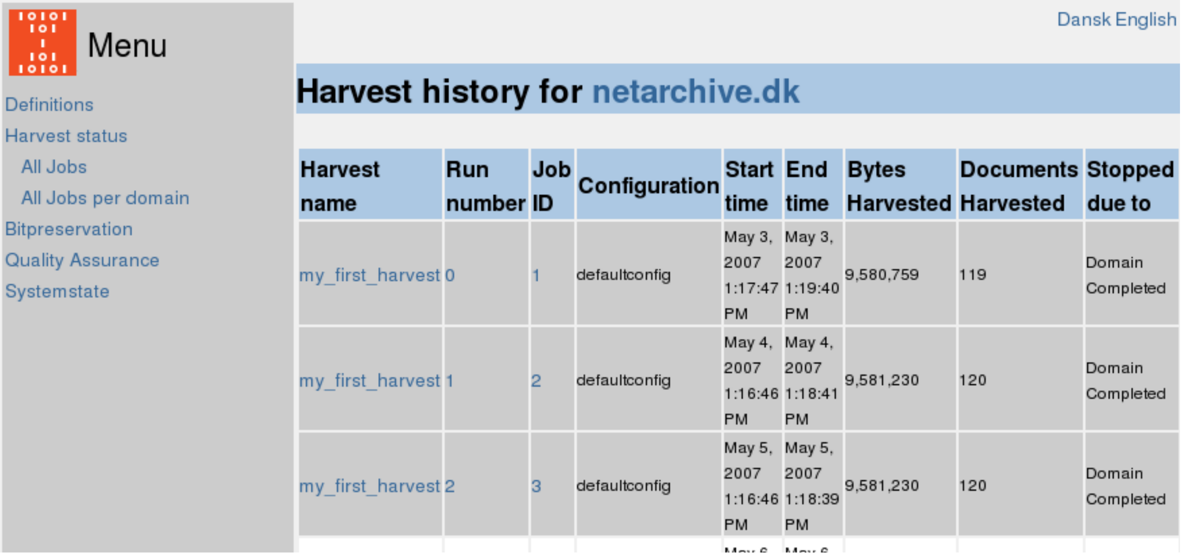
Run number	Start time	End time	Bytes Harvested	Documents Harvested	Total number of jobs	Number of failed jobs	N re
0	May 3, 2007 1:17:47 PM	May 3, 2007 1:19:40 PM	9,580,759	119	1 Show jobs	1	0
1	May 4, 2007 1:16:46 PM	May 4, 2007 1:18:41 PM	9,581,230	120	1 Show jobs	1	0
2	May 5, 2007 1:16:46 PM	May 5, 2007 1:18:39 PM	9,581,230	120	1 Show jobs	1	0
	May 6, 2007	May 6, 2007					

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 that failed and eventually got resubmitted.

History of a domain



Harvest history for netarchive.dk

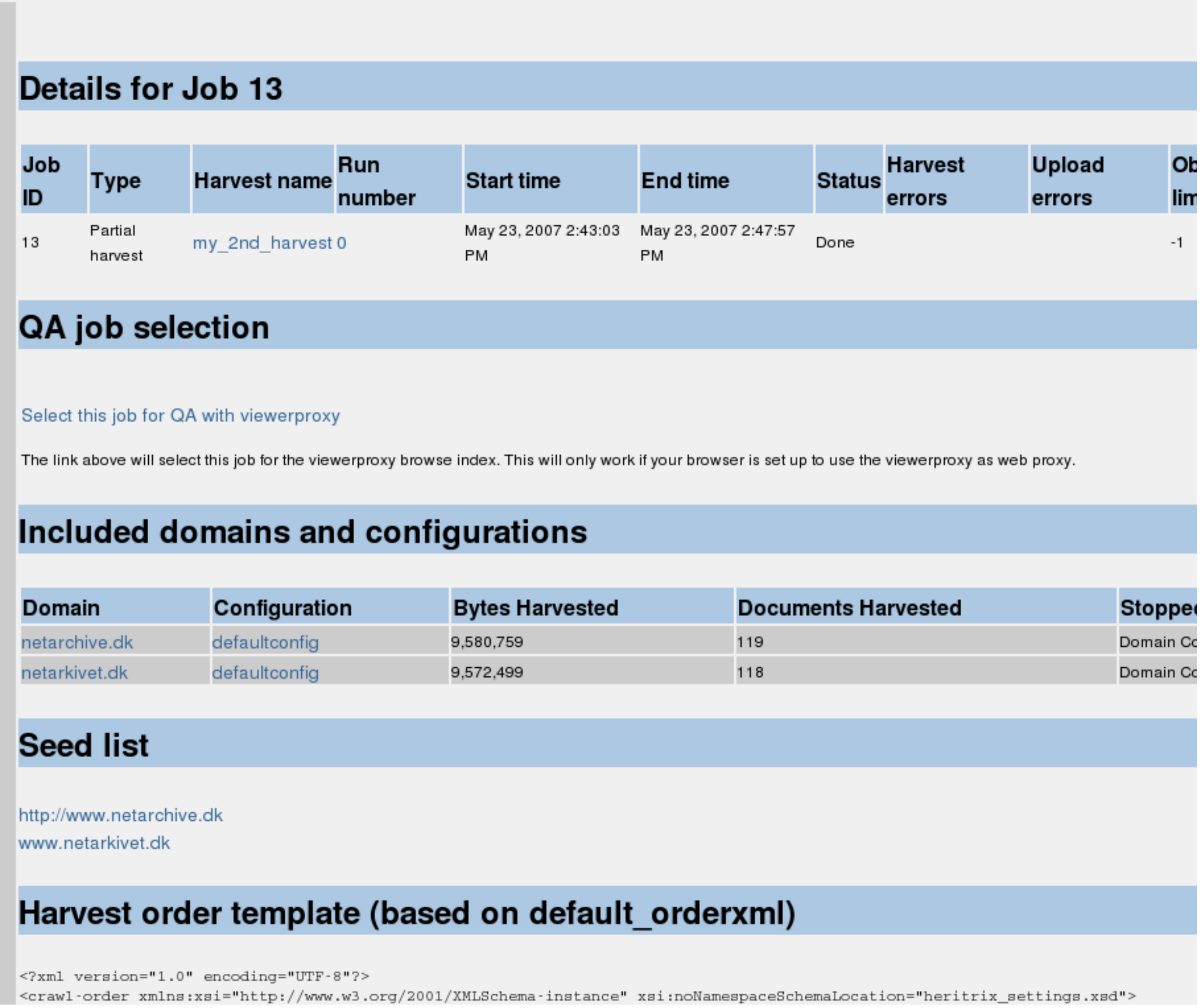
Harvest name	Run number	Job ID	Configuration	Start time	End time	Bytes Harvested	Documents Harvested	Stopped due to
my_first_harvest	0	1	defaultconfig	May 3, 2007 1:17:47 PM	May 3, 2007 1:19:40 PM	9,580,759	119	Domain Completed
my_first_harvest	1	2	defaultconfig	May 4, 2007 1:16:46 PM	May 4, 2007 1:18:41 PM	9,581,230	120	Domain Completed
my_first_harvest	2	3	defaultconfig	May 5, 2007 1:16:46 PM	May 5, 2007 1:18:39 PM	9,581,230	120	Domain Completed
				May 6,	May 6,			

If you want to see all the jobs connected to a specific domain, click on [***All jobs per domain***] *and search for the domain name.*

You will get a chronological list of the harvest definitions including the chosen domain.

This pages gives the same history information as the other two history pages and further more gives a "***Stopped due to***" information. This column will show the operator if a harvest was stopped unexpectedly of if the harvest hit the max-bytes limit for the chosen domain or if the harvest was stopped because of en error on the harvester machine.

Details on a job



Details for Job 13

Job ID	Type	Harvest name	Run number	Start time	End time	Status	Harvest errors	Upload errors	Ob lim
13	Partial harvest	my_2nd_harvest	0	May 23, 2007 2:43:03 PM	May 23, 2007 2:47:57 PM	Done			-1

QA job selection

Select this job for QA with viewerproxy

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

Included domains and configurations

Domain	Configuration	Bytes Harvested	Documents Harvested	Stoppe
netarchive.dk	defaultconfig	9,580,759	119	Domain Co
netarkivet.dk	defaultconfig	9,572,499	118	Domain Co

Seed list

http://www.netarchive.dk
www.netarkivet.dk

Harvest order template (based on default_orderxml)

```
<?xml version="1.0" encoding="UTF-8"?>
<crawl-order xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="heritrix_settings.xsd">
```

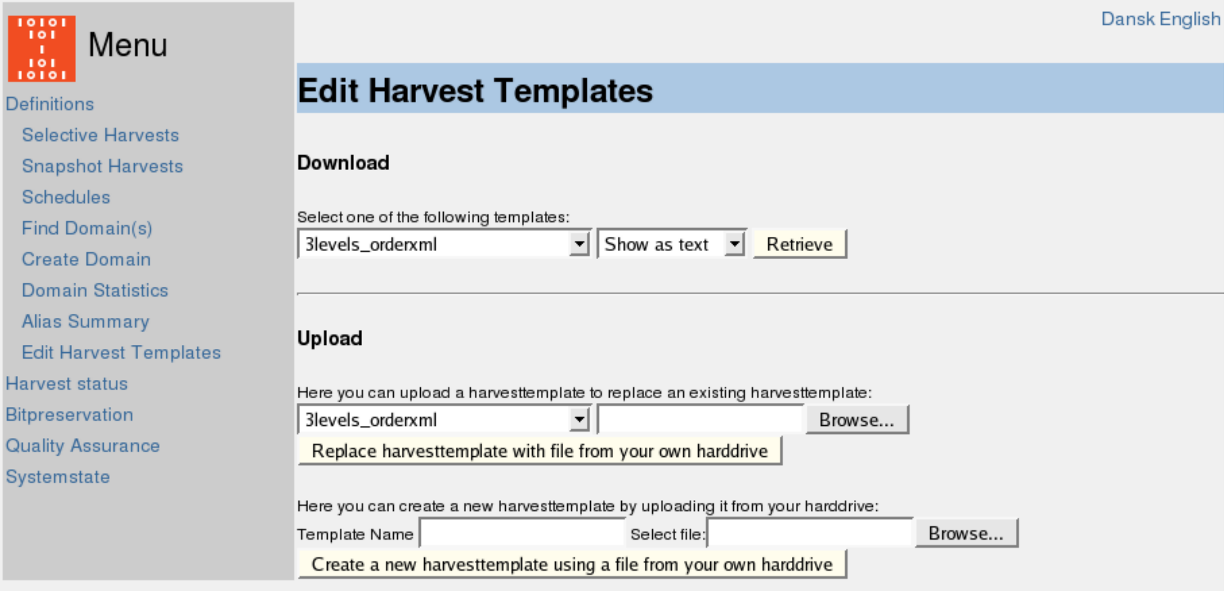
Clicking on a jobID on any of the harvest history pages will give you a very detailed report on the job.

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

Furthermore the page shows the complete seedlist used with the job and the complete “***Harvest order template***” as well as detailed error information in case of errors. The two latter is mainly for advanced users debugging specific crawls where things didn't go as expected.

edit

Harvester Templates



Edit Harvest Templates

Download

Select one of the following templates: 3levels_orderxml; Show as text

Upload

Here you can upload a harvesttemplate to replace an existing harvesttemplate: 3levels_orderxml

Here you can create a new harvesttemplate by uploading it from your harddrive:
Template Name · Select file:

The [***Edit Harvest Templates***] is used for managing the harvester templates. It enables you to both download and upload templates from/to the system database.

Download

The download part lets you view existing templates as either plain text or XML in the browser window or download existing templates to your local computer.

Select the template you want to view/download in the first select box, select the method in the second and press [***Retrieve***]

Upload

In the Upload section you can either update an existing template or create a new one.

To update an existing template, select the template to update in the first select box and browse for a file on your local computer with the *[**Browse**] button.*

By pressing [***Replace harvesttemplate with file from your own harddrive****] you will overwrite the chosen template in the database.*

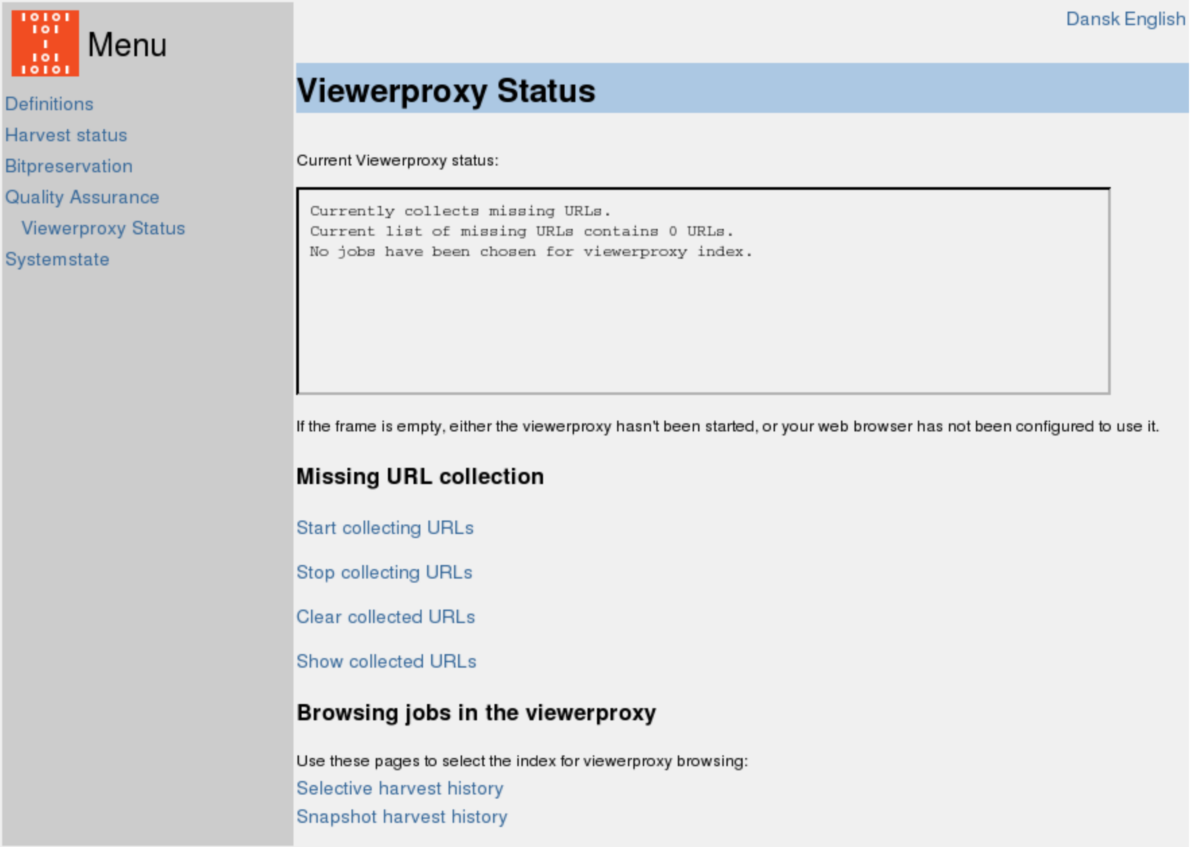
To create a new template give the new template a name in the “Template Name” box and select a file from your local computer with *[**Browse**].*

By pressing [***Create a new harvesttemplate using a file from your own harddrive***] you will add the new template with the given name to the database.

When using the upload functions the uploaded files will be checked against certain rules to ensure that the templates contain specific elements used by the NetarchiveSuite system.

edit

Quality Assurance



Viewerproxy Status

Current Viewerproxy status:

```
Currently collects missing URLs.
Current list of missing URLs contains 0 URLs.
No jobs have been chosen for viewerproxy index.
```

If the frame is empty, either the viewerproxy hasn't been started, or your web browser has not been configured to use it.

Missing URL collection

Browsing jobs in the viewerproxy

Use these pages to select the index for viewerproxy browsing:

Quality assurance is done by browsing the archive for selected domains. If something is missing on the pages the system can be set to automatically collect all the missing URL's for later transfer to the harvesting system. Before doing Quality Control you need to setup your browser to use a proxyserver (see Quick Start Manual)

It is suitable to investigate one domain at a time (unless several domains are included in the same website-complex).

[***Start collecting URL's***] *Hereby starts the collection of URL's. The **Current Viewerproxy status** textbox shows if the system collects URL's or not – and how many URL's are currently collected.*

[***Stop collecting URL's***] *Collection of URL's can be stopped at any time.*

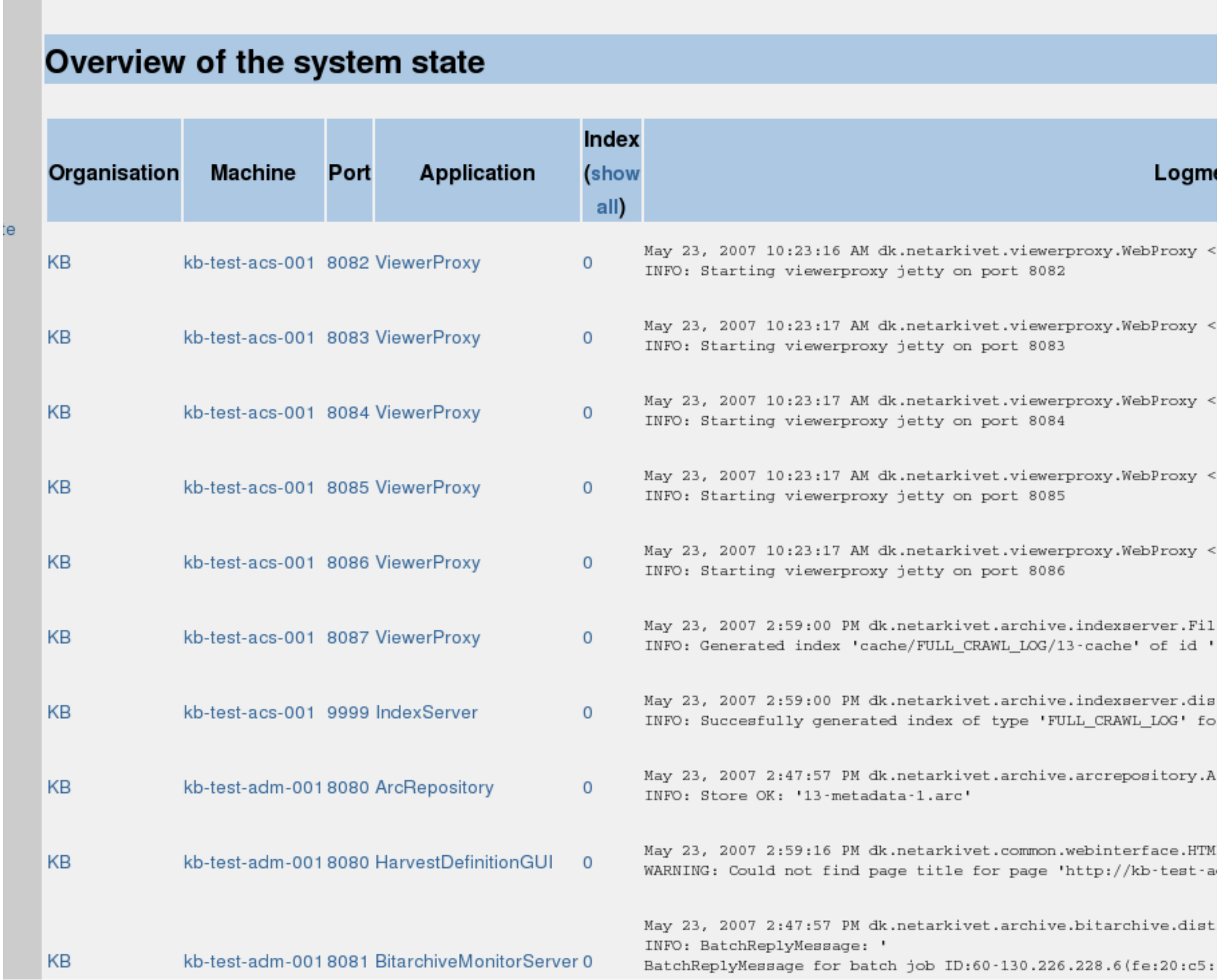
[***Clear collected URL's***] *The list of URL's can be cleared at any time e.g. when investigating a new domain starts. **NB!** This function can not be undone.*

[***Show collected URL's***] The list of URL's can be viewed at any time. The list can be copied and manually be added to relevant Seed lists for the relevant domains in the harvesting system.

edit

System State

The [***Systemstate***] pages lets the operator monitor the entire system (all machines and applications) from one central point.



Overview of the system state

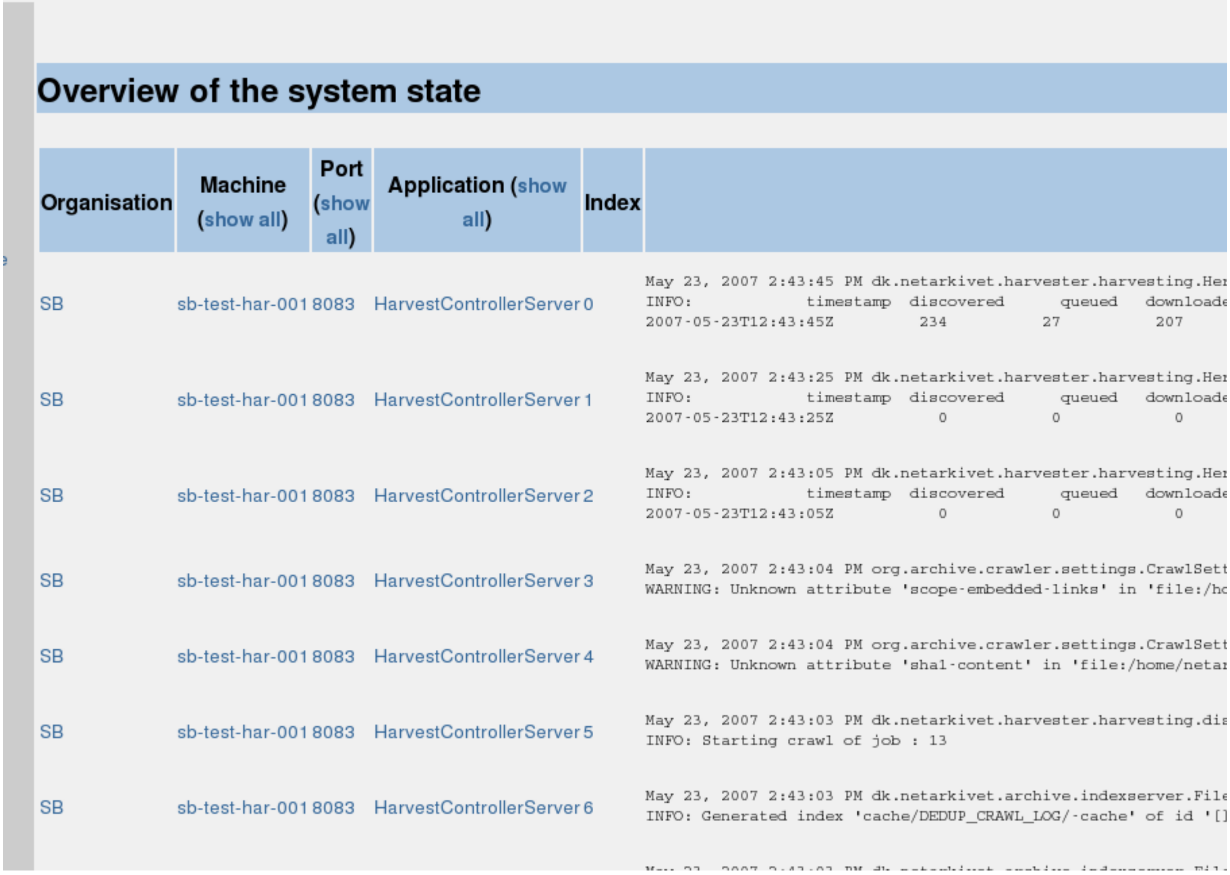
Organisation	Machine	Port	Application	Index (show all)	Logm
KB	kb-test-acs-001	8082	ViewerProxy	0	May 23, 2007 10:23:16 AM dk.netarkivet.viewerproxy.WebProxy < INFO: Starting viewerproxy jetty on port 8082
KB	kb-test-acs-001	8083	ViewerProxy	0	May 23, 2007 10:23:17 AM dk.netarkivet.viewerproxy.WebProxy < INFO: Starting viewerproxy jetty on port 8083
KB	kb-test-acs-001	8084	ViewerProxy	0	May 23, 2007 10:23:17 AM dk.netarkivet.viewerproxy.WebProxy < INFO: Starting viewerproxy jetty on port 8084
KB	kb-test-acs-001	8085	ViewerProxy	0	May 23, 2007 10:23:17 AM dk.netarkivet.viewerproxy.WebProxy < INFO: Starting viewerproxy jetty on port 8085
KB	kb-test-acs-001	8086	ViewerProxy	0	May 23, 2007 10:23:17 AM dk.netarkivet.viewerproxy.WebProxy < INFO: Starting viewerproxy jetty on port 8086
KB	kb-test-acs-001	8087	ViewerProxy	0	May 23, 2007 2:59:00 PM dk.netarkivet.archive.indexserver.Fil INFO: Generated index 'cache/FULL_CRAWL_LOG/13-cache' of id '
KB	kb-test-acs-001	9999	IndexServer	0	May 23, 2007 2:59:00 PM dk.netarkivet.archive.indexserver.dis INFO: Succesfully generated index of type 'FULL_CRAWL_LOG' fo
KB	kb-test-adm-001	8080	ArcRepository	0	May 23, 2007 2:47:57 PM dk.netarkivet.archive.arcrepository.A INFO: Store OK: '13-metadata-1.arc'
KB	kb-test-adm-001	8080	HarvestDefinitionGUI	0	May 23, 2007 2:59:16 PM dk.netarkivet.common.webinterface.HTM WARNING: Could not find page title for page 'http://kb-test-a
KB	kb-test-adm-001	8081	BitarchiveMonitorServer	0	May 23, 2007 2:47:57 PM dk.netarkivet.archive.bitarchive.dist INFO: BatchReplyMessage: ' BatchReplyMessage for batch job ID:60-130.226.228.6(fe:20:c5:

The initial view is the last log-message from every machine and every application.

This can be narrowed down to single "Organisations" / "Machines" / "Ports" / "Applications" and extended to not only show the last log-message but the last 100 log-messages (don't do that for the initial view of everything).

To narrow the view press either of the links on the page – when narrowed the view can be extended again with the [***Show all***] buttons that will dynamically appear in the headlines of the table.

The following is a view of one harvester instance on one specific machine (narrowed down by machine + port + application):



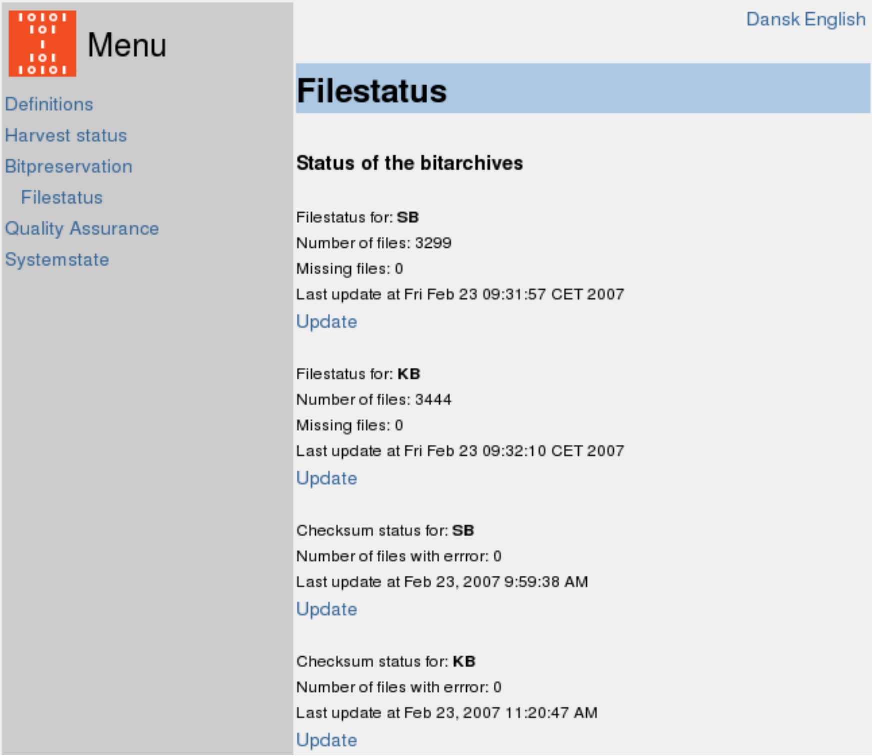
Overview of the system state

Organisation	Machine (show all)	Port (show all)	Application (show all)	Index	
SB	sb-test-har-001	8083	HarvestControllerServer	0	May 23, 2007 2:43:45 PM dk.netarkivet.harvester.harvesting.Her INFO: timestamp discovered queued downloade 2007-05-23T12:43:45Z 234 27 207
SB	sb-test-har-001	8083	HarvestControllerServer	1	May 23, 2007 2:43:25 PM dk.netarkivet.harvester.harvesting.Her INFO: timestamp discovered queued downloade 2007-05-23T12:43:25Z 0 0 0
SB	sb-test-har-001	8083	HarvestControllerServer	2	May 23, 2007 2:43:05 PM dk.netarkivet.harvester.harvesting.Her INFO: timestamp discovered queued downloade 2007-05-23T12:43:05Z 0 0 0
SB	sb-test-har-001	8083	HarvestControllerServer	3	May 23, 2007 2:43:04 PM org.archive.crawler.settings.CrawlSett WARNING: Unknown attribute 'scope-embedded-links' in 'file:/ho
SB	sb-test-har-001	8083	HarvestControllerServer	4	May 23, 2007 2:43:04 PM org.archive.crawler.settings.CrawlSett WARNING: Unknown attribute 'sha1-content' in 'file:/home/netar
SB	sb-test-har-001	8083	HarvestControllerServer	5	May 23, 2007 2:43:03 PM dk.netarkivet.harvester.harvesting.dis INFO: Starting crawl of job : 13
SB	sb-test-har-001	8083	HarvestControllerServer	6	May 23, 2007 2:43:03 PM dk.netarkivet.archive.indexserver.File INFO: Generated index 'cache/DEDUP_CRAWL_LOG/-cache' of id '[]

If you load this page just while a harvester instance is restarting you might get a JMX-error. The same thing will happen if one of the configured applications does not run or does not respond. So the system state will in some sense also discover non functional applications.

edit

Bit Preservation



Filestatus

Status of the bitarchives

Filestatus for: **SB**
Number of files: 3299
Missing files: 0
Last update at Fri Feb 23 09:31:57 CET 2007
Update

Filestatus for: **KB**
Number of files: 3444
Missing files: 0
Last update at Fri Feb 23 09:32:10 CET 2007
Update

Checksum status for: **SB**
Number of files with errror: 0
Last update at Feb 23, 2007 9:59:38 AM
Update

Checksum status for: **KB**
Number of files with errror: 0
Last update at Feb 23, 2007 11:20:47 AM
Update

The [***Bitpreservation***] interface lets you control active checks of the status of the underlying bitarchive. This only applies if your installation uses the NetarchiveSuite bitarchive application.

The interfaces lets you initiate two types of checks on every copy of the files in the archive: Filestatus and Checksum status.

In the example on the screen dump there are two bitarchive instances – **SB** and **KB**.

The [***Update***] buttons let you update the status for both files and checksums for both bitarchive instances. The page will give you the Filestatus as

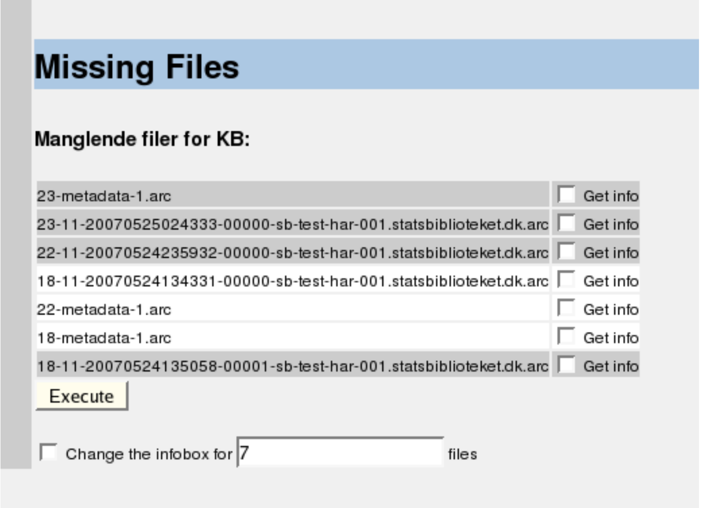
- Number of files
- Missing files
- Last updated at

and the Checksum status as

- Number of files with error
- Last updated at

The Filestatus checks are rather fast because only the existence of the files are checked whereas the Checksum status checks can take days/weeks for lager archives depending on the number of CPUs and the IO-speed of your hard drives.

Missing Files



Missing Files

Manglende filer for KB:

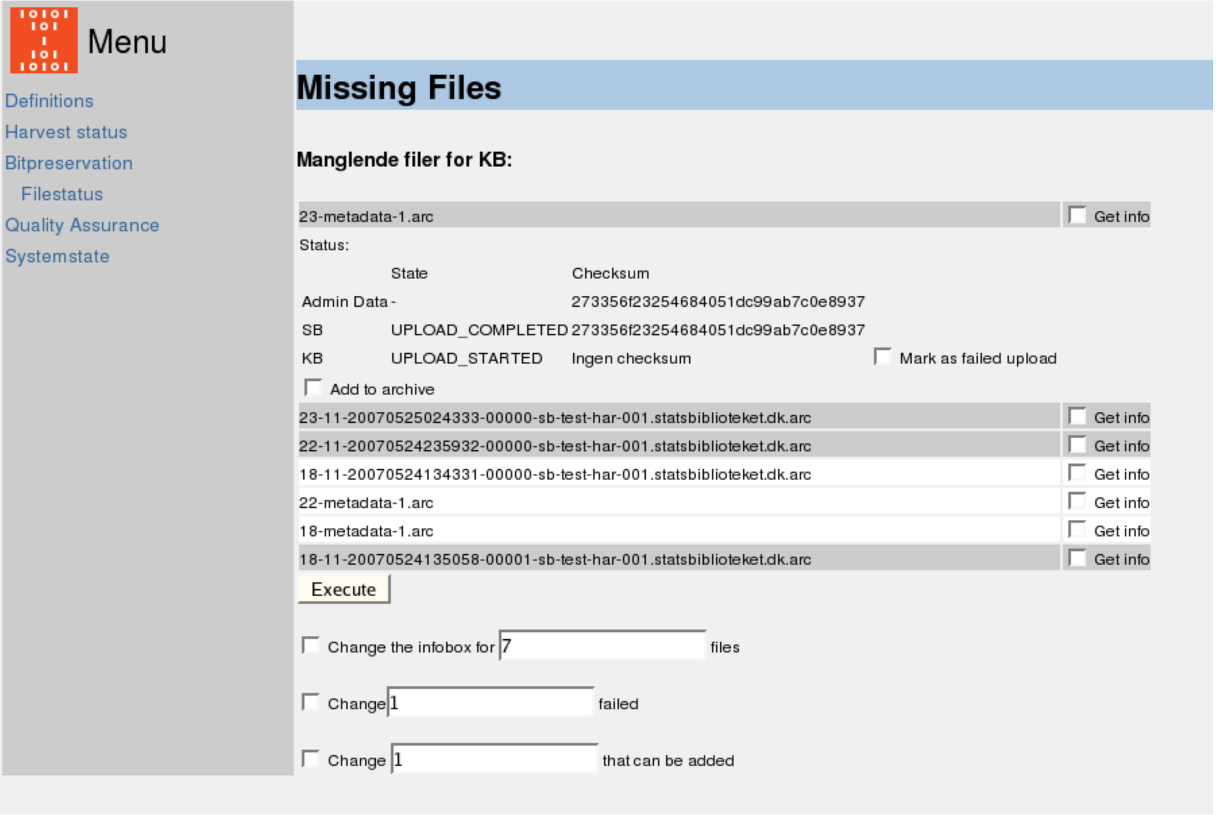
23-metadata-1.arc	□ Get info
23-11-20070525024333-00000-sb-test-har-001.statsbiblioteket.dk.arc	□ Get info
22-11-20070524235932-00000-sb-test-har-001.statsbiblioteket.dk.arc	□ Get info
18-11-20070524134331-00000-sb-test-har-001.statsbiblioteket.dk.arc	□ Get info
22-metadata-1.arc	□ Get info
18-metadata-1.arc	□ Get info
18-11-20070524135058-00001-sb-test-har-001.statsbiblioteket.dk.arc	□ Get info

□ Change the infobox for 7 files

If files are missing on one instance of the bitarchive a [***Show missing files****] button will appear right next to the line with the number of missing files*

For each missing file you can select “Get info”. With the “Change the infobox for”-field in the bottom of the screen you can select a number of files in one operation.

Pressing [***Execute****] makes the system get a fresh status on the files and there checksums from both copies (out of which one is missing) and from the administrative system so that there are always three checksums available for each file.*



Missing Files

Manglende filer for KB:

23-metadata-1.arc □ Get info

Status:

	State	Checksum	
Admin Data	-	273356f23254684051dc99ab7c0e8937	
SB	UPLOAD_COMPLETED	273356f23254684051dc99ab7c0e8937	
KB	UPLOAD_STARTED	Ingen checksum	□ Mark as failed upload

□ Add to archive

23-11-20070525024333-00000-sb-test-har-001.statsbiblioteket.dk.arc	□ Get info
22-11-20070524235932-00000-sb-test-har-001.statsbiblioteket.dk.arc	□ Get info
18-11-20070524134331-00000-sb-test-har-001.statsbiblioteket.dk.arc	□ Get info
22-metadata-1.arc	□ Get info
18-metadata-1.arc	□ Get info
18-11-20070524135058-00001-sb-test-har-001.statsbiblioteket.dk.arc	□ Get info

□ Change the infobox for 7 files

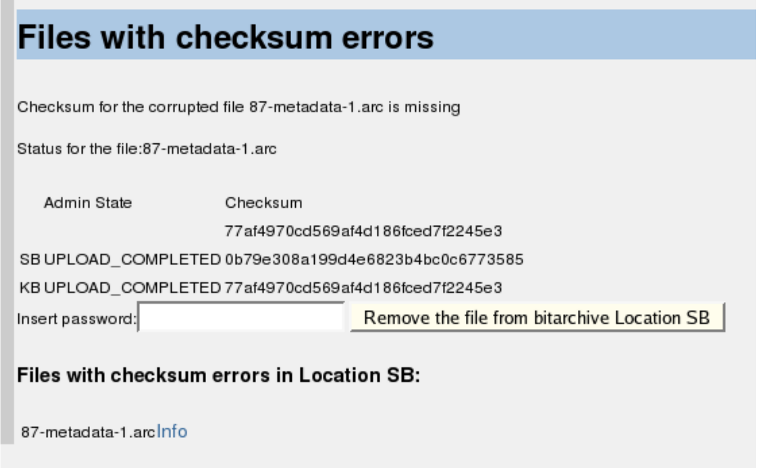
□ Change 1 failed

□ Change 1 that can be added

If the two “remaining” checksums (On the screen dump “273356f23254684051dc99ab7c0e8937”) are identical the system allows you to add the missing file to the bitarchive instance that had lost it.

This requires you to click “***Mark as failed upload****” and “*__Add to archive__” and then press [***Execute****]. Both marking files as failed uploads and marking them for addition can be done for a number of files in one operation. '*

Checksum Errors



Files with checksum errors

Checksum for the corrupted file 87-metadata-1.arc is missing

Status for the file:87-metadata-1.arc

	Admin State	Checksum
		77af4970cd569af4d186fced7f2245e3
SB	UPLOAD_COMPLETED	0b79e308a199d4e6823b4bc0c6773585
KB	UPLOAD_COMPLETED	77af4970cd569af4d186fced7f2245e3

Insert password:

Files with checksum errors in Location SB:

87-metadata-1.arcInfo

In case of a checksum error this error can be corrected through the interface. It requires two steps. First a deletion of the bad file and afterwards a re-upload of a good file with the above missing-files functionality.

To remove a bad file you need to type a security password and press [***Remove the file from bitarchive Location SB***]. The bad file will not be completely removed but moved to an 'attic' directory on the bitarchive server holding the bad file.

edit

User Manual (last edited 2008-06-12 16:24:39 by KaareChristiansen)