Quick Start Manual

This is a short primer to get a simple version of the NetarchiveSuite system up and running. Users with little technical knowledge can evaluate the software by installing it as described here.

It uses a pre-built script that starts all components on the same machine. This allows you to start experimenting with the functionality without having to do any more setup than absolutely necessary.

It should not require much technical skill to evaluate the system. What it does require is a computer running a Linux operating system and with sun java 1.6 or above installed. You do not need root/administrator access.

Going through this quick start should take about an hour.

Search manual

Download as pdf

Installation of the Quickstart system

We have prepared a bash shell script that starts all the necessary components on one machine. We will use this script throughout this quickstart manual to allow you to get a feel for what the system can do and how it works without having to deal with issues of distributing to other servers.

Base system required

For the quick startup, NetarchiveSuite requires:

- A Linux system with a minimum of 2GB free diskspace. (The minimum diskspace can be configured, but this is a reasonable minimum amount of space in which to store the harvested data.) Note that for the quickstart, you must be able to run a browser on the machine that you run the system on - this is an artifact of the quickstart system and is not the case in the full system. Root access is not required.
- Running Ant application - Java based build tool like make
- Sun Java SE (Standard Edition) JDK version 1.6.0_19 (or later) running on the Linux system.
  Newer versions of Sun Java 1.6 will probably work, but have not been tested. The latest download version of Sun Java 6 SE is "JDK 6 Update 24" (03 February 2011).

To check that you have the right version of Java do the following

- start a terminal login to the linux system as a ordinary user
- check java version is version 1.6.0_19 (or higher) by writing:

```
$ java -version
```

you should then see something like

```
linux>java -version
java version "1.6.0_19"
Java(TM) SE Runtime Environment (build 1.6.0_19-b04)
Java HotSpot(TM) Server VM (build 16.2-b04, mixed mode)
```

Downloading

Download of the newest release is described here

- Create a directory for the download e.g. directory

```
$ mkdir netarchive
```

- Download the relevant NetarchiveSuite zip and put it in the *netarchive* directory you created earlier.

Note: Instead of downloading a NetarchiveSuite.zip you can also build it yourself from the svn trunk:

```
$ svn export
https://sbforge.org/svn/netarchivesuite/trun
k .
$ cd trunk
$ ant releasezipball
$ mv NetarchiveSuite.zip
../NetarchiveSuite.zip
```

Setup JMS

NetarchiveSuite uses JMS for inter-process communication. JMS is the Java Messaging Service, which provides asynchronous communication between processes. You do not need any knowledge of JMS to use NetarchiveSuite. However you need to make sure that there are not already JMS brokers running on your system using PORT 7676.

Currently only the open-source version of Sun's JMS implementation is supported, since some functionality of other implementations does not match our assumptions well.

To download and install it, do the following:

- Open this link in a browser window http://mq.dev.java.net/downloads.html

- Click the Linux Link under version 4.4 Binary Downloads to download a file openmq4_4-installer-Linux_X86.zip (or later version)
- Save the download file to the *netarchive* directory you created earlier
- Goto the directory

```
$ cd ~/netarchive
```

- Unpack the zip file (this creates a directory openmq4_4-installer), and run the X-Windows installer. The installer ask you to choose an install-home (choose netarchive/MessageQueue), and a JDK.

```
$ unzip openmq4_4-installer-Linux_X86.zip
$ cd openmq4_4-installer; ./installer
```

- Set necessary environment variables: IMQ_HOME, IMQ_VARHOME, IMQ_ETCHOME)

```
$ export IMQ_HOME=$HOME/netarchive/MessageQueue/mq
$ export IMQ_VARHOME=$IMQ_HOME/var
$ export IMQ_ETCHOME=$IMQ_HOME/etc
```

- Run imqbroker in order to create settings file

```
$ chmod +x $IMQ_HOME/bin/imqbrokerd
$ $IMQ_HOME/bin/imqbrokerd
```

- check that

```
imqbrokerd
```

  starts and that the last message is

```
"Broker <localhost>:7676 ready"
```

- stop the imqbroker by pressing

```
control-C
```

- edit settings to allow for enough listeners to a queue by doing
  edit

```
$IMQ_VARHOME/instances/imqbroker/props/config.properties
```

- uncomment and specify count=20 for listeners by changing line

```
#            imq.autocreate.queue.maxNumActiveConsumers
```

  to

```
imq.autocreate.queue.maxNumActiveConsumers=20
```

To start it, do the following:

```
$ cd netarchive
$ $IMQ_HOME/bin/imqbrokerd &
```

Installation

Download following files to the *netarchive* directory:

- RunNetarchiveSuite.sh.
- deploy_standalone_example.xml.

The first script is a simple script for doing all the steps during deployment. It takes a NetarchiveSuite package ('.zip'), a configuration file (the second file), and a temporary installation directory as arguments (in the given order). The different ports used by the application for communication are included in the deploy_standalone_example.xml file.

In the configuration file all the applications are placed on one machine, e.g. the current machine (localhost)

When the installation script is run it will unpack the installation files into the *netarchive/deploy* directory and install NetarchiveSuite into the */home/test/QUICKSTART* directory (using ssh). It assumes that user 'test' already exists. Remember to check, that a Sun JVM is in the path for the test (instead of GNU java compiler, that is default with some Linux'es.). If you already have a Quickstart installation, the existing bitarchive, database and admin.data files will be untouched. You must explicit remove any previous installation, if you want a clean empty installation.

```
$ chmod +x RunNetarchiveSuite.sh
$ ./RunNetarchiveSuite.sh
NetarchiveSuite.zip
deploy_standalone_example.xml deploy/
```

Note that if you have not setup your automatic ssh test user login (using key based login), you need to login some times before the installation finish successfully. You must also have permission to ssh and scp to test@localhost (try e.g *ssh test@localhost*)

The script creates a deployment folder named "QUICKSTART" in e.g. /home/test/QUICKSTART, which contains methods for starting and stopping NetarchiveSuite, and starts the whole NetarchiveSuite. The files to run the installation will be placed in the ~/netarchive/deploy dir

- Start a web browser. Note that it is important that the browser is started on the same machine as the simple harvest script is run on
- Setup the browser to proxy on port 8070 and exclude localhost and the hostname (used by the Heritrix GUI) e.g. in firefox:

```
Choose in the firefox toolbar:
Edit->Preferences->Advanced->Network->Settin
gs
Checkmark:
Manual Proxy Configuration
and add:
Proxy: localhost
Port: 8070
No Proxy for: localhost
```

- Write following url in the started browser http://localhost:8074/HarvestDefinition
- You can now see the webinterface in the browser. You can now create, run and browse according to the following or the User Manual
- You can stop and start the entire NAS system with:

```
ssh test@localhost
cd QUICKSTART
./conf/killall.sh
./conf/startall.sh
```

- If you want to try other deploy examples, then go to "Examples of deploy configuration files" in the Installation Manual.

Running a simple harvest

Walkthrough of the definition and execution of a simple harvest.

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

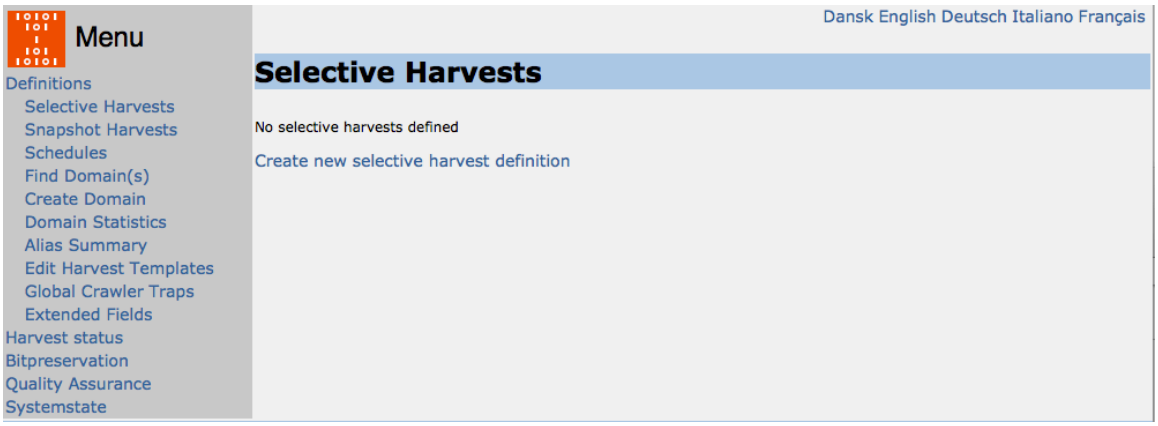
Setting up the harvest

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

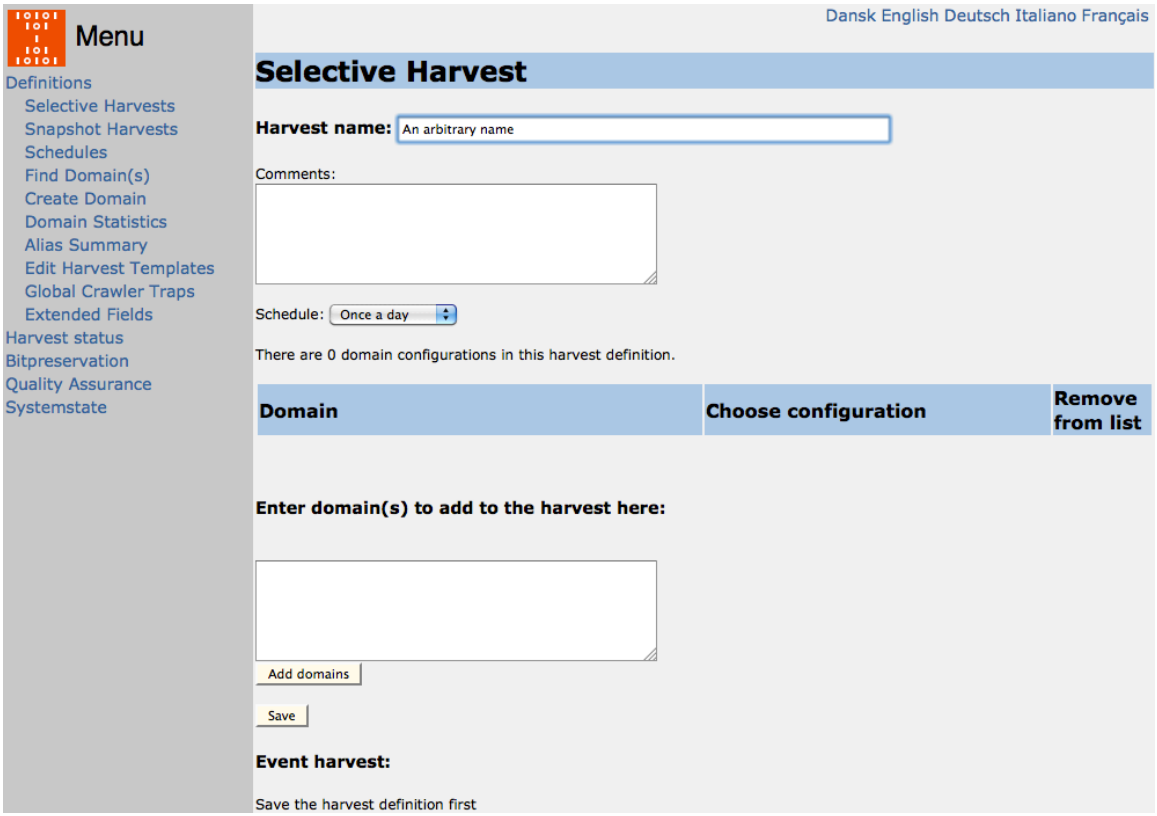
Open http://localhost:8074/HarvestDefinition in a browser on the local machine.

You can now define a new harvest.

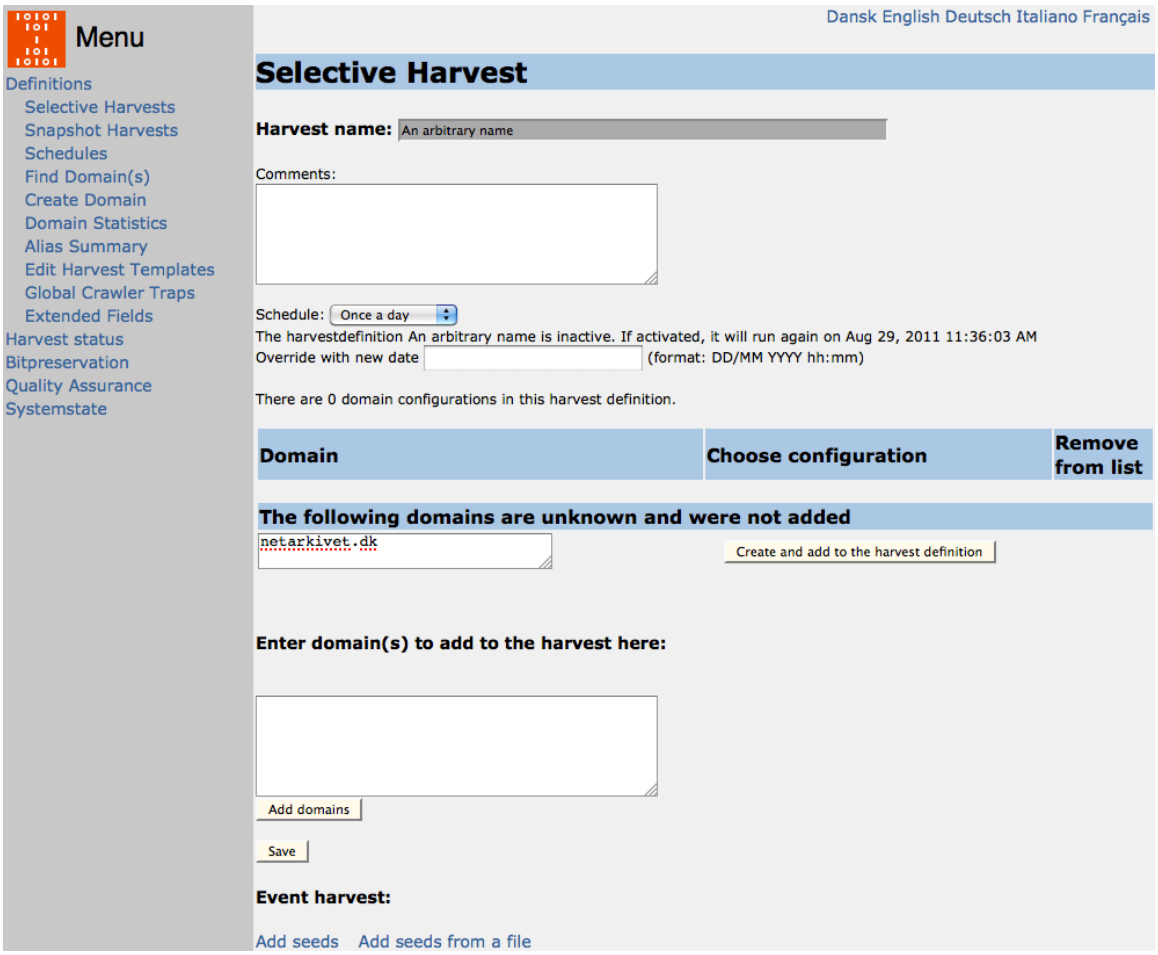
Click 'Selective Harvests' under menu 'Definitions'



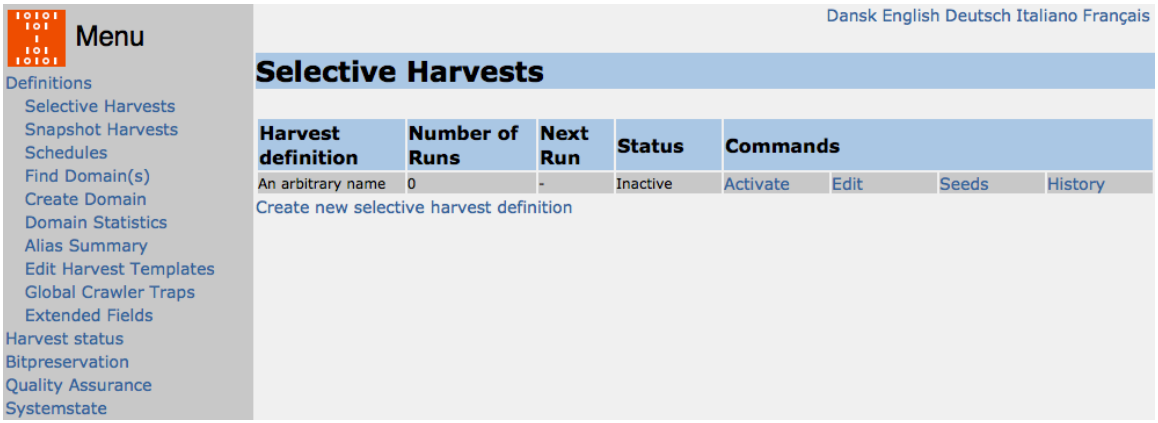
Click 'Create new harvest definition' under the (empty) table of existing harvests.



Enter an arbitrary name for the harvest in the top. Enter some second-level domain name (e.g., netarchive.dk) in the box and press 'Add domains'. 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.



Since the domain didn't exist in the database, the system suggests you add it. Click 'Create and add to harvest definition'. You can now click 'Save' on the 'Selective Harvest' page



Now you have defined a harvest definition for this domain. It will however not start a harvest before it is changed to active state.

Click 'Activate' for the newly defined harvest.

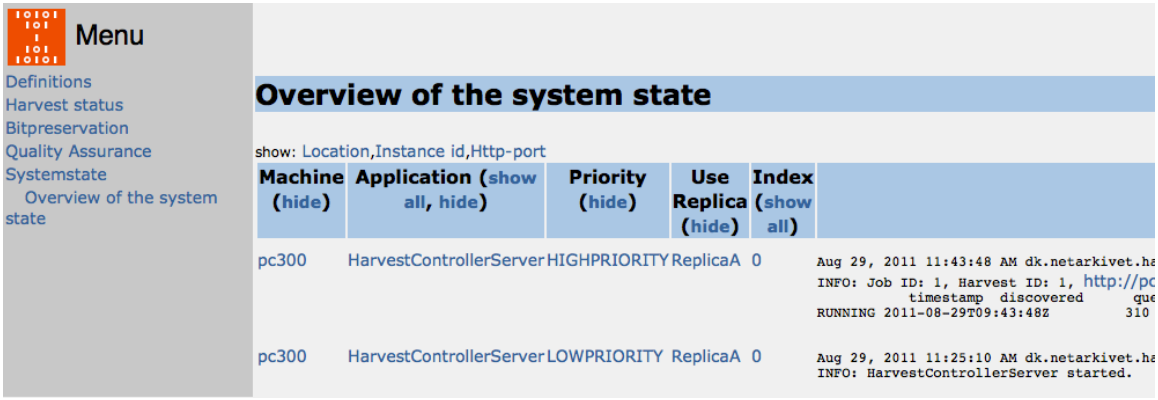
The harvest definition will generate harvest jobs.

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.

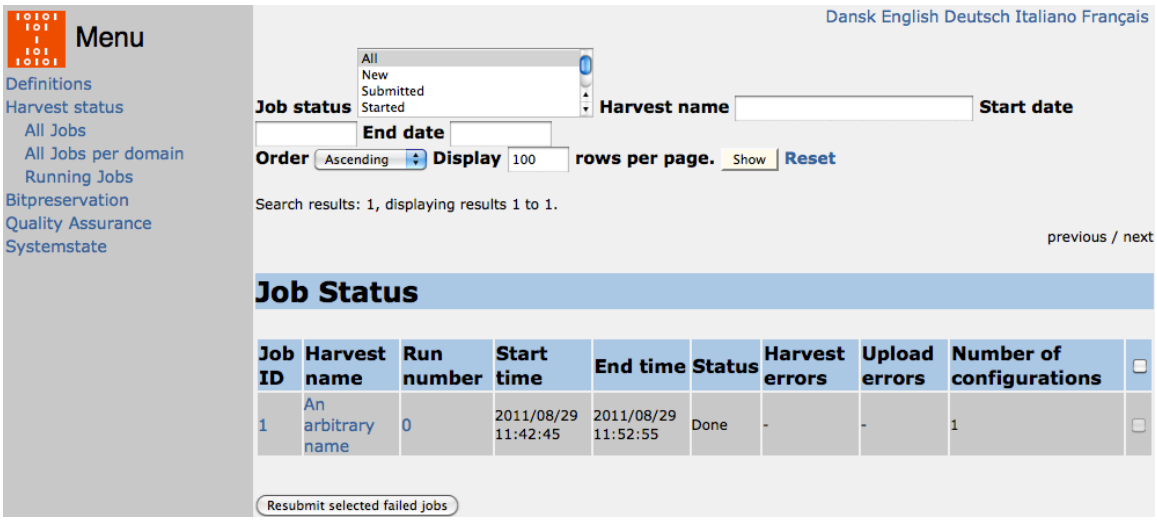
Now you can monitor the system state for what is going on in the various components. That way you can see how the harvester is proceeding with the job:

Go to the System Status page by clicking 'Systemstate'. Click on the application HarvestControllerServer. 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.



Use the System Status and Job Status pages to monitor your job. You can also jump to the Heritrix GUI by clicking on the log line URL e.g. Harvest ID: 1 pc300:8192 as long as the job is running by using the std. 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. Alternatively, just replace the PC name with "localhost" in the URL, e.g. http://localhost:8192.)

Go to the Job status page by clicking 'Harvest status'. Set wanted jobs status to 'All' and click 'Show'. It will take a little while for the job to finish and to upload the harvested files to the !NetarchiveSuite archive (about 5 min.). Refresh the page until the 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

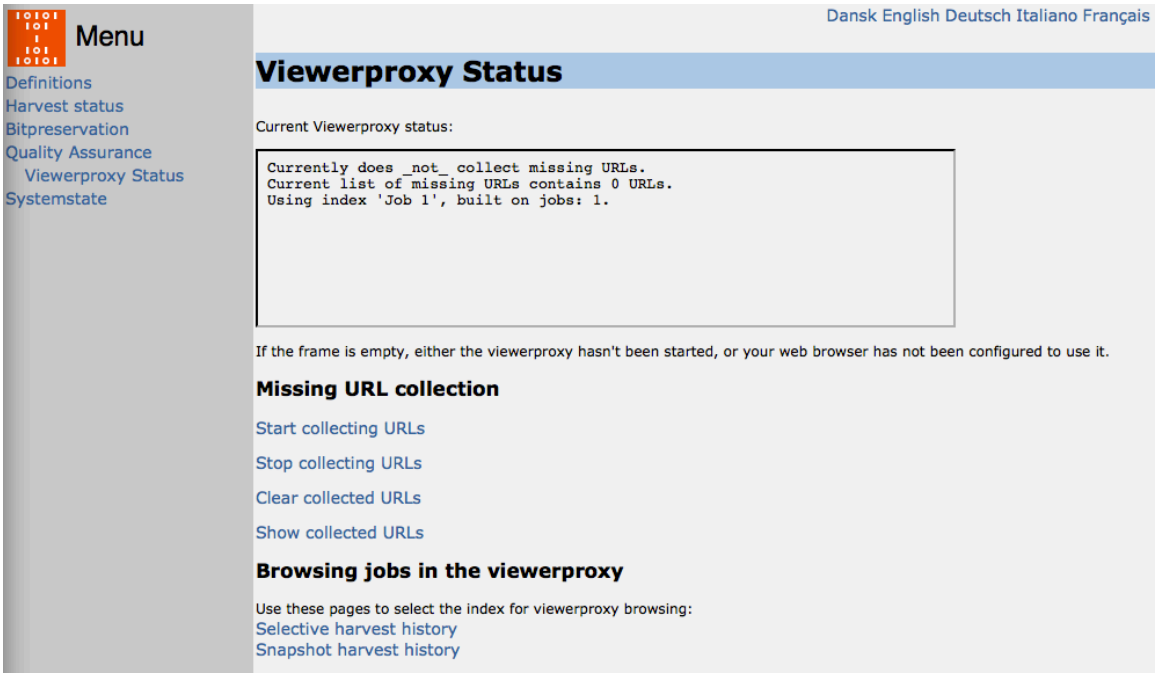
Once that some web pages have been harvested, you can use the viewerproxy part to view them.

Before it is ready, it needs to know which material you wish to browse.

Go to the 'Harvest Status' page, select to show 'All' jobs and click 'Show'. Click on the link with the Job Id.

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.



Now simply enter the URL that you started harvesting from (with www), e.g. *www.netarchive.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, when you are doing Quality Assurance on the harvested material.

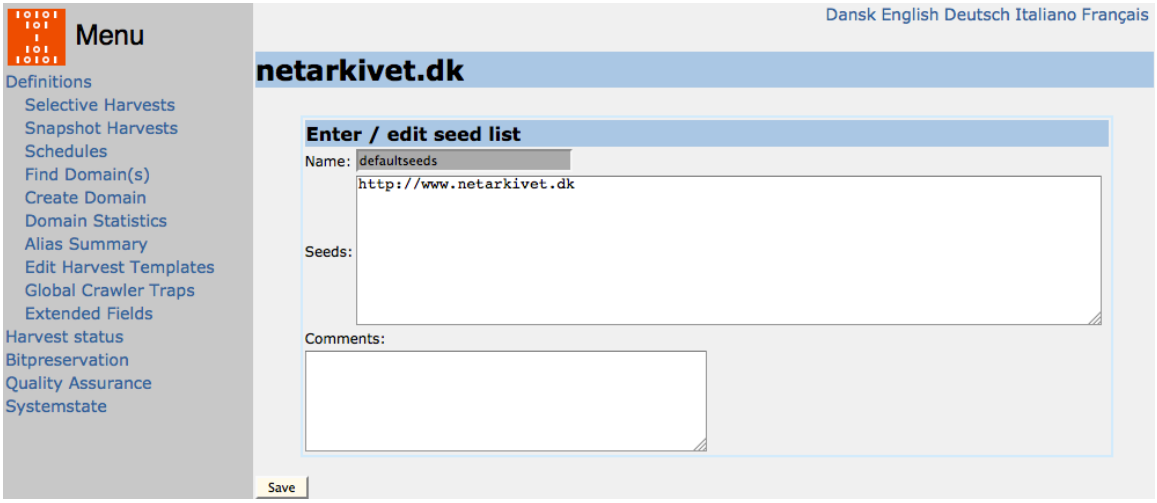
To try this, go back to the viewerproxy status page and click 'Start collecting URLs'. Now browse in the collected material until you find a page or image that did not get harvested. Go back to the viewerproxy status page and click 'Show collected URLs'.

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

Let us make sure these URLs are harvested next time we harvest the domain.

Copy the URLs for your harvested domain that were found missing into the clipboard. Go to the domain definition page by clicking 'Find Domain(s)' under 'Definitions' and search for your domain.

You will now get a page with information used when harvesting that domain. In this case, we wish to add the collected URLs to the list of seeds we start our web harvests from.



On the domain definition page, click 'Edit' next to the seed list.

Add the URLs from the clipboard to the seed list and press 'Update'.

These URLs will be used as seeds the next time the domain is harvested, i.e. the harvest will include these URLs in the harvest. To see this in effect, create another harvest of this domain following all the steps above. Wait for the domain to finish harvesting, then go to the 'job status' page for the new job. Limit the viewerproxy to the new job only and browse the material again. The URLs that were missing last time should now be found.

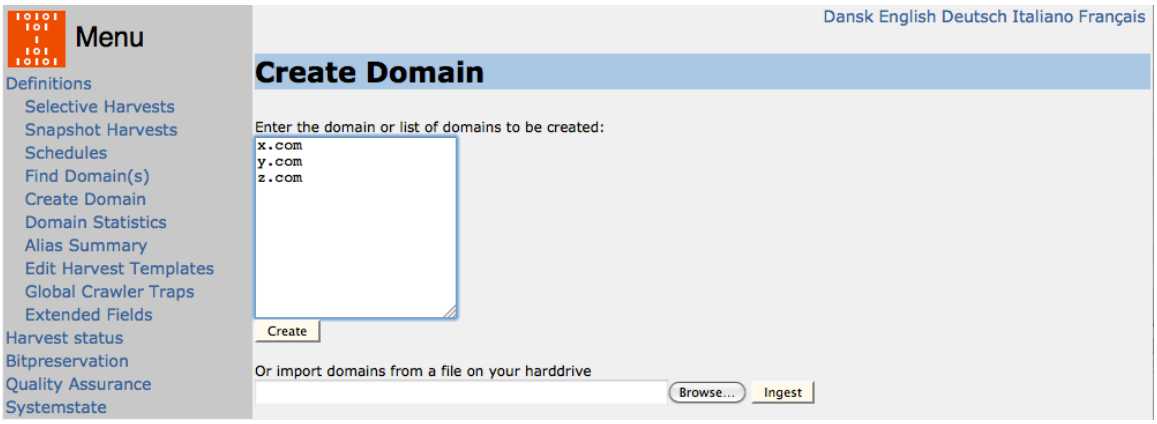
Running a snapshot harvest

A snapshot harvest harvests all known domains up to a given byte limit, i.e. a limit of bytes that you harvest from each domain, i.e. a limit of bytes that you harvest from each domain. This is used for nationwide harvests of "'all'" domains. You can also use "Max number of objects per domain" ("-1" means without limit). The best practice is to use byte limits or object limits - not a combination.

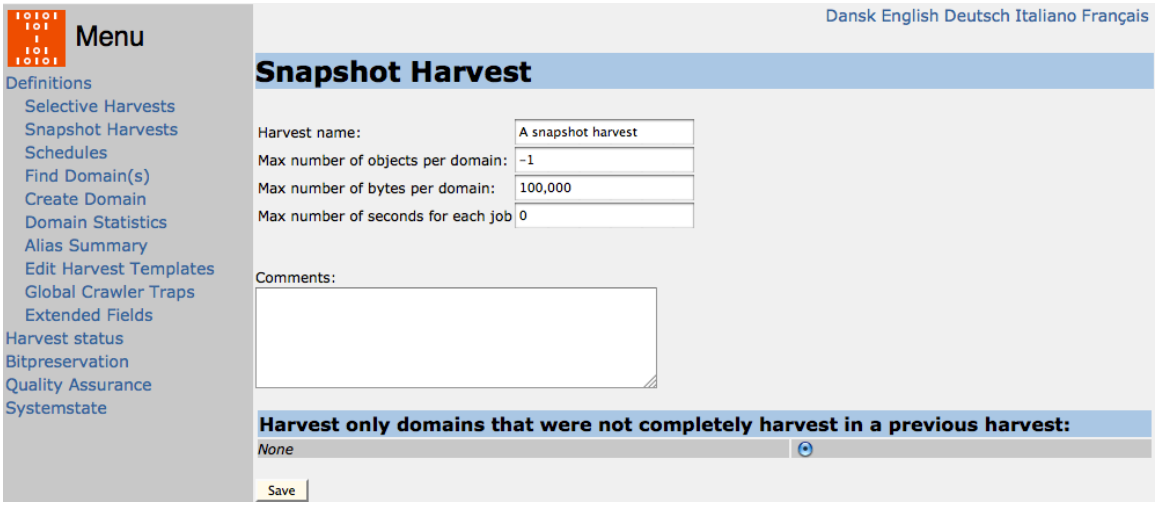
Each domain has one "default configuration" automatically generated when the domain is created. The default configuration is used to determine how to harvest the domain in a snapshot harvest. Typically, the default configuration is good enough for most purposes, but if you want to have a domain excluded from the snapshot harvest (e.g. if the domain is outside the group you're interested in) you may want to set the harvest limit on the default configuration for that domain to 0. The default configuration is also the one used in a selective harvest unless another configuration is chosen in the drop-down menu on the selective harvest page. The other way to control how a snapshot harvest is executed is by choosing a different harvest template. Descriptions of how harvest templates work are in the user manual.

NetarchiveSuite has support for mass creation of domains, for instance by ingesting (loading) a list of domains given by a national TLD (top-level-domain) administrator.

To ingest, go to the "Create Domain" page under "Definitions" and specify the file containing the list of domains. You can also type domains in the text window, but this is only usable for a smaller number of domains. The list should be a newline-separated list of domain names including the top level domain, but not including subdomains, protocol specifications or URL paths. Thus *netarkivet.dk* or *archive.org* are useable, while *http://foo.com*, *bar.dk/hest* or *news.bbc.co.uk* are not. What is considered a top-level domain is configurable. Typically it would be a country top level domain for most countries (like .dk, .fr etc), but fore some special cases it makes more sense to define the top level a little further down (for instance .co.uk). See how to configure this in the [Installation Manual 3.16]. When the file is specified, press "Ingest" and wait while the domains are ingested. For a first test, you probably want to keep it to a fairly small number of sites, to make sure the test harvest doesn't take too long.



After ingest, you can click on 'Domain statistics' under 'Definitions' to see an overview of how many domains are registered under the TLDs. To create a snapshot definition, go to 'Snapshot harvests' and press 'Create new snapshot harvest'. The harvest definition presented will require you to enter a harvest name, and also allows you to add comments or changing the limit of how many bytes or objects to collect per domain. Keep this to a fairly low number for a first test, to make sure the harvest doesn't run too long.



When you have entered the information, press 'Save' and then press 'Activate'.

You can monitor the harvest and browse the harvested material exactly as you did in the previous harvests.

It is possible - only while the job is running - to access the Heritrix user interface on the harvester (See further details above or in the User Manual).

Next steps...

This concludes the quickstart manual. While you can of course continue to play around with this simple setup, there are numerous options and possibilities that are not mentioned herein that are useful for scalability and for adapting the harvesting to your needs. Further information about installation and configuration can be found in the Installation Manual and more details on how to use the web interface can be found in the User Manual.

Enjoy!