

NetarchiveSuite Quick Start Manual

Printer friendly version

Contents	
1.	Introduction
2.	System overview
3.	Download and installation
1.	Base system required
2.	Downloading
3.	JMS
4.	Configuration
4.	Running a simple harvest
1.	Setting up the harvest
2.	Viewing the results
5.	Running a snapshot harvest
6.	Carrying on...

Introduction

edit

This manual provides instructions for quickly getting a basic NetarchiveSuite system up and running. 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.

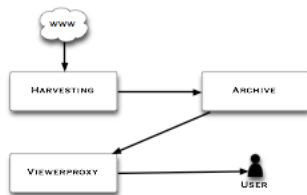
System overview

edit

The primary function of the NetarchiveSuite is to plan, schedule and archive web harvests of parts of the internet. We use Heritrix as our webcrawler. The NetarchiveSuite can organize three different kinds of harvests:

- Event harvesting (organize harvests of a set of domains related to a specific event, e.g. 9/11, Royal Weddings, and Elections).
- Selective harvesting (recurrent harvests of a set of domains).
- Snapshot harvesting (organizing a complete snapshot of all known domains)

The NetarchiveSuite is split into three main modules corresponding to harvesting, archiving and accessing via viewerproxy.



Please refer to the overview description for more details.

Download and installation

edit

For a quick start, 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.
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 running on the Linux system.
Newer versions of Java may work, but have not been tested. Older versions of Java will **not** work correctly. The latest download version of Sun Java 6 SE is "JDK 6 Update 19" (09 March 2010).

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

- start a terminal login to the linux system as an 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

- start a terminal login to the linux system as an ordinary user in a bash shell
- make a directory for the download e.g. directory `~/netarchive`

```
$ mkdir ~/netarchive
```
- start a web browser, e.g. Firefox
- follow the registration and download instructions on Get NetarchiveSuite and save the download file to the directory you created earlier
- there should now be a `NetarchiveSuite*.zip` file in the installation directory, e.g. `~/netarchive`

Note: Instead of downloading a NetarchiveSuite.zip you can also build itself from the SVN trunk:

```
$ svn checkout --username developername https://gforge.statsbiblioteket.dk
/svn/netarchivesuite/trunk
$ cd trunk
$ ant releasezipball
```

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 8100.

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 <https://mq.dev.java.net/downloads.html>
- click the Linux Link under version 4.1 Binary Downloads to download a file `mq4_1-binary-Linux_X86-20070816.jar` (or later)
- save the download file to the created directory you created earlier e.g. `~/netarchive`
- go to the directory

```
$ cd ~/netarchive
```
- unpack the jar file (this creates a directory `mq` and three files with licensing information)

```
$ jar xvf mq4_1-binary-Linux_X86-20070816.jar
```

- run imqbroker in order to create settings file


```
$ chmod +x ./mq/bin/imqbrokerd
```

 Set the IMQ_JAVAHOME environment variable to point to your java installation directory, e.g. /usr/java/jdk1.6.0_19:


```
$ export IMQ_JAVAHOME=/usr/java/jdk1.6.0_19
$ ./mq/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 ~/netarchive/mq/var/instances/imqbroker/props/config.properties
```

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


```
"#           imq.autocreate.queue.maxNumActiveConsumers"
->
"           imq.autocreate.queue.maxNumActiveConsumers=20"
```

To start it, do the following:

```
$ cd ~/netarchive
$ ./mq/bin/imqbrokerd &
```

Configuration

Assuming a releasezipball of NetarchiveSuite `NetarchiveSuite*.zip` is available in the directory `~/netarchive`, you must do the following to configure the NetarchiveSuite for your system:

- Download following attached files to e.g. `/home/test/netarchive`:

📄 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).

In the configuration file all the applications are placed on one machine (e.g. the current machine, `localhost`). If run directly it is run from the deploy directory `/home/test/netarchive/USER` and installed in e.g. `localhost:/home/test/USER`. It assumes, that you want to run this as user 'test'. So you need to have 'test' user on the current machine. If installation user is different from the 'test' user, remember to check, that a Sun JVM is in the path (instead of GNU java compiler, that is default with some Linux'es.). If you already have a USER installation, then remember also, that the existing bitarchive, database and admin.data files will be untouched. You must explicit remove any previous USER installation, if you want a clean empty installation.

E.g. (you should use "USER" as the installation name to make things easy)

```
cd /home/test/netarchive
bash RunNetarchiveSuite.sh NetarchiveSuite.zip deploy_standalone_example.xml USER/
#if you have not setup your ssh keygen correctly, you need to login some times before the installation finish
successfully. You must also have permission to ssh and scp to localhost ( try e.g "ssh localhost" and "scp
somefile localhost:")
```

The script creates a deployment folder named "USER" in e.g. `/home/test/netarchive`, which contains methods for starting and stopping NetarchiveSuite, and starts the whole NetarchiveSuite. It deploys the installation locally to `/home/test/USER`

- start a web browser by e.g. `$ firefox` 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->Settings
Checkmark:
Manual Proxy Configuration
and add:
Proxy: localhost
Port: 8070
No Proxy for: localhost, kb-test-way-001.kb.dk
```

- 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
- if you want to stop and start the entire NAS system, then

```
cd /home/test/netarchive/USER
./killall_NATIONAL_LIBRARY.sh
./startall_NATIONAL_LIBRARY.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

edit

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'

The screenshot shows the NetarchiveSuite web interface. On the left is a 'Menu' sidebar with a list of options: Definitions, Selective Harvests, Snapshot Harvests, Schedules, Find Domain(s), Create Domain, Domain Statistics, Alias Summary, Edit Harvest Templates, Global Crawler Traps, Harvest status, Bitpreservation, Quality Assurance, and Systemstate. The main content area is titled 'Selective Harvests' and contains the text 'No selective harvests defined' and a link 'Create new selective harvest definition'. At the top right of the interface, there are language options: Dansk, English, Deutsch, Italiano, Français. At the bottom, a status bar reads 'NetarchiveSuite Version: 3.12.0 status RELEASE, QUICKSTART'.

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

 **Menu**Dansk English Deutsch Italiano Français

Definitions

- Selective Harvests
- Snapshot Harvests
- Schedules
- Find Domain(s)
- Create Domain
- Domain Statistics
- Alias Summary
- Edit Harvest Templates
- Global Crawler Traps

Harvest status

- Bitpreservation
- Quality Assurance
- Systemstate

Selective Harvest

Harvest name:

Comments:


Schedule: ▾

Domain	Choose configuration	Remove from list
<p>Enter domain(s) to add to the harvest here:</p> <input style="width: 100%; height: 40px;" type="text"/>		

Event harvest:
Save the harvest definition first

NetarchiveSuite Version: 3.12.0 status RELEASE, QUICKSTART

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. You can add more domains if you want by repeating the procedure, but in this example we will only add one domain.



Menu

- Definitions
- Selective Harvests
- Snapshot Harvests
- Schedules
- Find Domain(s)
- Create Domain
- Domain Statistics
- Alias Summary
- Edit Harvest Templates
- Global Crawler Traps
- Harvest status
- Bitpreservation
- Quality Assurance
- Systemstate

Dansk English Deutsch Italiano Français

Selective Harvest

Harvest name:

Comments:

Schedule:

The harvestdefinition An arbitrary name is inactive. If activated, it will run again on Apr 21, 2010 10:30:12 AM
 Override with new date (format: DD/MM YYYY hh:mm)


Domain	Choose configuration	Remove from list
The following domains are unknown and were not added		
<input style="width: 100%;" type="text" value="netarkivet.dk"/>	<input type="button" value="Create and add to the harvest definition"/>	

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

Event harvest:
[Add seeds](#) [Add seeds from a file](#)

NetarchiveSuite Version: 3.12.0 status RELEASE, QUICKSTART

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



Menu

- Definitions
- Selective Harvests
- Snapshot Harvests
- Schedules
- Find Domain(s)
- Create Domain
- Domain Statistics
- Alias Summary
- Edit Harvest Templates
- Global Crawler Traps
- Harvest status
- Bitpreservation
- Quality Assurance
- Systemstate

Dansk English Deutsch Italiano Français

Selective Harvests

Harvest definition	Number of Runs	Next Run	Status	Commands			
An arbitrary name	0	-	Inactive	Activate	Edit	Seeds	History
Create new selective harvest definition							

NetarchiveSuite Version: 3.12.0 status RELEASE, QUICKSTART

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.

Dansk English Deutsch Italiano Français

Menu
 Definitions
 Harvest status
 All Jobs
 All Jobs per domain
 Bitpreservation
 Quality Assurance
 Systemstate

Only show job status in order

Job Status

Job ID	Harvest name	Run number	Status	Harvest errors	Upload errors	Number of configurations
1	An arbitrary name	0	New	-	-	1

NetarchiveSuite Version: 3.12.0 status RELEASE. QUICKSTART

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.

Menu
 Definitions
 Harvest status
 Bitpreservation
 Quality Assurance
 Systemstate
 Overview of the system state

Overview of the system state

show: Location,Machine,Instance id,Http-port,Priority,Use Replica

Application (show all, hide)	Index (show all)
HarvestControllerServer 0	Apr 21, 2010 10:38:09 AM dk.netarkivet.harvester.harvesting.HeritrixLau INFO: Job ID: 1, Harvest ID: 1, http://KB-TEST-WAY-001.kb.dk:8192 RUNNING timestamp discovered queued downloaded 2010-04-21T08:38:09Z 276 125 151 4(3)
HarvestControllerServer 0	Apr 21, 2010 10:12:24 AM dk.netarkivet.harvester.harvesting.distribute. INFO: HarvestControllerServer started.

NetarchiveSuite Version: 3.12.0 status RELEASE. QUICKSTART

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 KB-TEST-WAY-001.kb.dk:8192 as long as the job is running by using the std. Heritrix login "admin" and Password "adminPassword"

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".

Dansk English Deutsch Italiano Français

Menu

- Definitions
- Harvest status
- All Jobs
- All Jobs per domain
- Bitpreservation
- Quality Assurance
- Systemstate

Only show job status in order

Job Status

Job ID	Harvest name	Run number	Status	Harvest errors	Upload errors	Number of configurations
1	An arbitrary name	0	Done	-	-	1

NetarchiveSuite Version: 3.12.0 status RELEASE, QUICKSTART

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.

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.

Dansk English Deutsch Italiano Français

Menu

- Definitions
- Harvest status
- Bitpreservation
- Quality Assurance
- Viewerproxy Status
- Systemstate

Viewerproxy Status

Current Viewerproxy status:

Currently does not collect 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

- [Start collecting URLs](#)
- [Stop collecting URLs](#)
- [Clear collected URLs](#)
- [Show collected URLs](#)

Browsing jobs in the viewerproxy

Use these pages to select the index for viewerproxy browsing:

- [Selective harvest history](#)
- [Snapshot harvest history](#)

NetarchiveSuite Version: 3.12.0 status RELEASE, QUICKSTART

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'.

Menu

- Definitions
- Harvest status
- Bitpreservation
- Quality Assurance
 - Viewerproxy Status
- Systemstate

Dansk English Deutsch Italiano Français

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

- [Start collecting URLs](#)
- [Stop collecting URLs](#)
- [Clear collected URLs](#)
- [Show collected URLs](#)

Browsing jobs in the viewerproxy

Use these pages to select the index for viewerproxy browsing:

- [Selective harvest history](#)
- [Snapshot harvest history](#)

NetarchiveSuite Version: 3.12.0 status RELEASE, QUICKSTART

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.

The screenshot shows the 'netarkivet.dk' interface. On the left is a 'Menu' with options like 'Definitions', 'Selective Harvests', 'Snapshot Harvests', 'Schedules', 'Find Domain(s)', 'Create Domain', 'Domain Statistics', 'Alias Summary', 'Edit Harvest Templates', 'Global Crawler Traps', 'Harvest status', 'Bitpreservation', 'Quality Assurance', and 'Systemstate'. The main area is titled 'Enter / edit seed list'. It contains a 'Name' field with 'defaultseeds', a 'Seeds' text area with 'http://www.netarkivet.dk', and a 'Comments' text area. A 'Save' button is at the bottom left. The footer reads 'NetarchiveSuite Version: 3.12.0 status RELEASE. QUICKSTART'.

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.

edit

Running a snapshot harvest

edit

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. This is used for national-wide harvests of **all** domains. You can also use "Max number of objects per domain" ("-1" means without limit). Best praxis 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 for 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. 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.

Dansk English Deutsch Italiano Français

Menu

- Definitions
 - Selective Harvests
 - Snapshot Harvests
 - Schedules
 - Find Domain(s)
 - Create Domain
 - Domain Statistics
 - Alias Summary
 - Edit Harvest Templates
 - Global Crawler Traps
- Harvest status
- Bitpreservation
- Quality Assurance
- Systemstate

Create Domain

Enter the domain or list of domains to be created:

x.com
y.com
z.com

Create

Or import domains from a file on your harddrive

Browse... Ingest

NetarchiveSuite Version: 3.12.0 status RELEASE, QUICKSTART

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.

Dansk English Deutsch Italiano Français

Menu

- Definitions
 - Selective Harvests
 - Snapshot Harvests
 - Schedules
 - Find Domain(s)
 - Create Domain
 - Domain Statistics
 - Alias Summary
 - Edit Harvest Templates
 - Global Crawler Traps
- Harvest status
- Bitpreservation
- Quality Assurance
- Systemstate

Snapshot Harvest

Harvest name: A snapshot harvest

Max number of objects per domain: -1

Max number of bytes per domain: 100,000

Comments:

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

None

Save

NetarchiveSuite Version: 3.12.0 status RELEASE, QUICKSTART

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).

Carrying on...

edit

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. The Development tab has information on how to program new or altered functionality for NetarchiveSuite in Java.

Enjoy!

