

# NetarchiveSuite Installation Manual

Printer friendly version

## Contents

1. Introduction
  2. Configuration basics
  3. Choose an installation scenario
  4. Choose a type of database
  5. Choose a JMS broker
  6. Choose the set of machines taking part in the installation/deployment
    1. Configure monitoring (allocating JMX and RMI ports)
  7. Additional configurations
    1. Configure notifications
    2. Select a file datatransfer method
    3. Configure job-generation
    4. Configure Domain Granularity
    5. Configure Heritrix process
    6. Configure web page look
    7. Configure security
      1. Core classes
      2. Third-party classes
  8. The actual deployment of the NetarchiveSuite
    1. Standard commandline settings
      1. The CLASSPATH
      2. Logging
      3. JMX settings
      4. Select the appropriate settings.file for the application
      5. JVM options
    2. Admin machine
      1. Starting the GUIApplication
      2. Starting the BitarchiveMonitorApplication instances
    3. Harvester machines
    4. Bitarchive machines
    5. Access servers
  9. Starting and stopping the NetarchiveSuite
    1. NetarchiveSuite application startup order
    2. NetarchiveSuite application stopping order
  10. Monitoring an running instance of NetarchiveSuite
- Appendix A : Installing external software
    1. Installing and configuring a JMS broker
      1. Obtaining JMS
      2. Installing the JMS broker
      3. Configuring the JMS broker
      4. Starting and stopping JMS
        1. How to empty queues
    2. Installing and configuring FTP

- 1. Starting and stopping a Proftpd server
- Appendix B : Configurable settings in the NetarchiveSuite
- Appendix C : Plugins in the NetarchiveSuite
- Appendix D : Managing Heritrix harvest templates (order.xml)
  1. Mandatory elements in the NetarchiveSuite and their role
    1. 1) The QuotaEnforcer
    2. 2) The DeDuplicator
    3. 3) The "http-headers" element
    4. 4) The Archiver element
    5. 5) The Scope element
      1. The anatomy of a decidingscope
        1. A. The header
        2. B. The defining deciderule
        3. C. Standard harvest rules
        4. D. Define general crawlertraps to be avoided
    2. The HarvestTemplateApplication tool
    3. Predefined harvest templates
      1. Templates w/ DomainScope
      2. Templates w/ HostScope
      3. Templates w/ PathScope
- Appendix E : Migrate the heritrix templates to NetarchiveSuite 3.6.0+
  1. How to convert from the former scopes to a decidingscope

## Introduction

edit

This manual describes how to install and configure the NetarchiveSuite web archive software package. It includes description of how to obtain and install required libraries, how to install the software on separate machines, what command line options and configuration file changes are necessary, and how to start the programs. It then goes on to explain the other parameters available for tuning the behaviour of NetarchiveSuite. It does not explain how to extend the functionality of the system (see the Developer Manual for this) or how to use the running system (see the User Manual: for this).

The intended audience of this manual is system administrators who will be responsible for the actual installation and setup of NetarchiveSuite as well as technical personnel responsible for proper operation of NetarchiveSuite. Knowledge of Unix system administration is expected, and some familiarity with XML and Java is an advantage.

## Configuration basics

edit

Almost all configuration of NetarchiveSuite applications is done by changing the `settings.xml` file, which by default is located in the `conf` directory. This file is organized in

a hierarchy with settings for the various modules and applications having their own parts. All settings can also be set from the command line by specifying the `-D` option with a dotted path of the XML path. For instance, using the command line option `-Dsettings.common.tempDir=/tmp` corresponds to changing the XML file to read (in part)

```
<settings>
  <common>
    <!-- Common temporary directory for all applications. -->
    <tempDir>/tmp</tempDir>
  ...

```

Using command line options override whatever is set in the `settings.xml` file, but allows only one value per setting. In the `settings.xml` file, you can specify several values by repeating a section. For instance, the settings for "language" and "siteSection".

Whether to use a `settings.xml` file per application or per machine or just have the same file for all machines is up to you. In this manual, we will assume a shared `settings.xml` file for all machines and define the settings that vary using command-line options. Depending on how you choose to perform your installation, you may want to have a different approach.

## Choose an installation scenario

edit

NetarchiveSuite can be installed in a number of different ways, with varying numbers of machines on different sites. There is a number of separate applications in play, most of which can be put on separate machines as needed. To keep clear what is necessary for which setups, we will consider the following types of setup:

- **A. Single-machine setup.** This corresponds to the setup used in the Quick Start Manual, where all applications run on the same machine, and file transfer can be done simple by copying files locally. It is the simplest setup, but does not scale very well. Note that the scripts used in Quick Start Manual resets the system at every restart, including deleting all harvested material. Obviously, this is not the intent for a running installation, so those scripts cannot be used in production environments as they are.
- **B. Single-site setup.** In this scenario, multiple machines are involved, necessitating file transfer between machines and multiple installations of the code. However, the machines are expected to be within the same firewall, so port setup should be no problem.
- **C. Single-site setup with duplicate archive.** This expands on the single-site setup in that more than one copy of the archives are used, using the concept of separate "Locations" to indicate the duplicates.
- **D. Multi-site setup.** When more than one site is involved, separated by firewalls, extra issues of opening ports and specifying the correct site come into play. This is the most complex scenario, but also the more secure against systematic errors, hacking, and other disasters.

Setups C and D involves having a distributed bitarchive. In these setups we have the the bitarchive distributed on at least two Locations, here called LocationA and LocationB. One of them must be designated as the Location that by default executes batch jobs -- typically the

Location with the greater amount of processing power. These Location informations must be written to the general settings.xml before deployment:

```
<arcrepository>
  ...
  <!-- The names of all bit archive locations in the
        environment, e.g., "LocationA" and "LocationB". -->
  <location>
    <name>LocationA</name>
  </location>
  <location>
    <name>LocationB</name>
  </location>
  <!-- Default bit archive to use for batch jobs (if none is specified)
  -->
  <batchLocation>LocationA</batchLocation>
</arcrepository>
```

## Choose a type of database

edit

The NetarchiveSuite can use three types of database:

- embedded Derby database (default)
- external Derby database
- MySQL database

By default, the NetarchiveSuite uses an embedded Derby. If you choose this option, you only have to do this before you launch the NetarchiveSuite applications:

```
cd <installationdir>/harvestdefinitionbasedir
unzip fullhddb.jar
```

If you want to use an external Derby, you have to do the following

- start Derby separately:
  - cd "directory with the extracted database" (e.g. <installationdir>/harvestdefinitionbasedir)
  - export CLASSPATH=<installationdir>/lib/db/derby-net-10.1.1.0.jar:<installationdir>/lib/db/derby-10.1.1.0.jar
  - java org.apache.derby.drda.NetworkServerControl start [-p port]

The default port is 1527.

For the NetarchiveSuite to use this external database, you

- set the database.specificsclass to `dk.netarkivet.harvester.datamodel.DerbyServerSpecifics`
- set the database.url to `jdbc:derby://localhost:1527/fullhddb` (substitute the server host for localhost, and 1527 for correct port)
- need to add a permission to the policy file used by your installation, if you use security. The following will allow NetarchiveSuite to access a Derby database on port 1527:

```
grant {
    permission java.net.SocketPermission "127.0.0.1:1527",
        "connect, resolve";
};
```

**Firewall note:** You will need to allow the GUIApplication and the HarvestTemplateApplication to be able to access port 1527 on the server where you run the database.

More details on using Derby as a server are available on [the derby pages](#).

If you want to use a MySQL database, you have to

- set the database.specificsclass to `dk.netarkivet.harvester.datamodel.MySQLSpecifics`
- set the database.url correctly: `jdbc:mysql://localhost/fullhddb?user=root&password=secret` (substitute the server host for localhost, and
- Install the MySQL database (v. 5.0.X) on a machine of your choice
- Download a `mysql-connector-java-5.0.X-bin.jar` from <http://dev.mysql.com/downloads/connector/j/5.0.html>
- add a permission to the policy file used by your installation, if you use security. The following will allow NetarchiveSuite to access MySQL on localhost on the default port 3306.

```
grant {
    permission java.net.SocketPermission "127.0.0.1:3306",
        "connect, resolve";
};
```

**Firewall note:** You will need to allow the GUIApplication and the HarvestTemplateApplication to be able to access port 3306 on the server where you run the database.

This jar must then be added to the classpath for the applications, that accesses the database: GUIApplication and HarvestTemplateApplication

You can do this manually, when starting these applications. Alternatively, you can add the `mysql-connector-java-5.0.X-bin.jar` to the `lib/db` directory, and modify `build.xml` accordingly:

- Add a line `"db/mysql-connector-java-5.0.X-bin.jar"` to the property 'jarclasspath' just below the line `"db/derby-10.1.1.0.jar"`.
- Add a line `"<include name="db/mysql-connector-java-5.0.X-bin.jar"/>"` below `<include name="db/derby-10.1.1.0.jar"/>`

You can then generate a new NetarchiveSuite zipball with `"ant zipball"`.

This assumes, that you have downloaded the source distribution of the NetarchiveSuite.

## Choose a JMS broker

edit

The NetarchiveSuite requires the use of a JMS broker. The installation and startup of a JMSbroker is described in Appendix A. In the below extract of conf/settings.xml, the JMSbroker resides at machine1.domain, and listens for messages on port 7676. You must also select a JMS environmentName. This allows you have more than one running installation of the NetarchiveSuite, each with its own environmentName, and makes it easy to cleanup the JMS queues associated with a given environmentName. The NetarchiveSuite currently only supports one kind of JMS broker, so only the 'broker', 'port', and 'environmentName' can be changed.

```
<jms>
  <!-- Selects the broker class to be used. Must be a subclass of
        dk.netarkivet.common.distribute.JMSConnection.-->
  <class>dk.netarkivet.common.distribute.JMSConnectionSunMQ</class>
  <!-- The JMS broker host contacted by the JMS connection -->
  <broker>localhost</broker>
  <!-- The port the JMS connection should use -->
  <port>7676</port>
  <!-- The name of the environment in which this code is running, e.g.
        PROD, RELEASETEST, NHC,... Common prefix to all JMS
channels
-->
  <environmentName>PROD</environmentName>
</jms>
```

**Firewall note:** The machine that runs the JMS broker must be accessible from all machines in the installation on not only port 7676, but also port 33700 (from RMI).

## Choose the set of machines taking part in the installation/deployment

edit

When you have chosen your setup, you must decide on the number of machines, you want to use in the deployment of the NetarchiveSuite. For setup A, the answer is of course one. For the setup B-D, the answer is more complicated.

At the Danish installation, we operate with 4 kinds of machines:

- Admin machine (one server): Here we deploy one or more BitarchiveMonitorApplications (one for each Location), one ArcrepositoryApplication, and one GUIApplication (which also controls scheduling). The latter application is the only application using a database.
- harvester machines (one or more): Here we deploy the HarvesterControllerApplications and their associated SideKicks.
- bitarchive machines (one or more): These machines only run one BitarchiveApplication each (there must be at least one for each location).
- access servers (one or more): On these machines, we have the ViewerproxyApplication enabling us to browse in already stored webpages, and the IndexServerApplication. The latter must only be installed on one of the access-servers, as there can only be one in the system,

Apart from the HarvestControllerApplications and their associated SideKicks, there is no requirement that the applications are placed like this, but we will use it as an example throughout the rest of the manual. In the standard setup used in our test-environment, we have 9 machines:

```
1 bitarchive server (on Location A)
2 bitarchive servers (on Location B)

1 admin machine (placed on Location A)
2 harvester-machines (placed on Location A)
2 harvester-machines (placed on Location B)

1 access server (placed on Location A)
```

## Configure monitoring (allocating JMX and RMI ports)

Monitoring the deployed NetarchiveSuite relies on JMX (Java Management Extensions). Each application in the NetarchiveSuite needs its own JMX-port and associated RMI-port, so they can be monitored from the NetarchiveSuite GUI, and using *jconsole* (see below). You need to select a range for the JMX-ports. In the example below, the chosen JMX/RMI-range begins at 8100.

On each machine you need to set the JMX and RMI ports, using the settings `settings.common.jmx.port` and `settings.common.jmx.rmiPort`.

**Firewall Note:** This requires that the admin-machine has access to each machine taking part in the deployment on ports 8100-8300.

You need to select a password for the JMX monitorRole, and replace the string "JMX\_MONITOR\_ROLE\_PASSWORD\_PLACEHOLDER" with the selected password in two files: the `conf/jmxremote.password`, and the settings file used. When launching the applications we define the path to the JMX passwordfile on the commandline:

```
-Dsettings.common.jmx.passwordFile=INSTALLATION_DIR/conf/jmxremote.password
```

The applications will automatically register themselves for monitoring at the GUI application, if the `StatusSiteSection` is deployed.

For the `StatusSiteSection` to work, your logging must be configured to use `java.util.logging` with the `dk.netarkivet.monitor.logging.CachingLogHandler` enabled, see *Logging* under *Command Line Options* below

## Additional configurations

edit

### Configure notifications

NetarchiveSuite can send notifications of serious system warnings or failures to the system-owner by email. This is implemented using the Notifications plugin, see Appendix C. Several settings in the `settings.xml` must be changed for this to work:

The setting `settings.common.notifications.receiver` (recipient of notifications), `settings.common.notifications.sender` (the official sender of the email, and receiver of any bounces), and `settings.common.mail.server` (the proper mail-server to use):

```
<common>
  ...
```

```

<notifications>
  <!-- Which class to instantiate to handle error notifications -->
  <class>dk.netarkivet.common.utils.EmailNotifications</class>
  <!-- The receiver of emails -->
  <receiver>example@netarkivet.dk</receiver>
  <!-- The stated sender of emails (and receiver of bounces)-->
  <sender>example@netarkivet.dk</sender>
</notifications>
<!-- Settings for sending email. Currently mail is only used for email
notifications. -->
<mail>
  <!-- The email server to use -->
  <server>examplesmtpserver.netarkivet.dk</server>
</mail>

```

Alternatively, the class `dk.netarkivet.common.utils.PrintNotifications` can be used. This will simply print the notifications to `stderr` on the terminal.

```

<common>
  ...
  <notifications>
    <!-- Which class to instantiate to handle error notifications -->
    <class>dk.netarkivet.common.utils.PrintNotifications</class>
  </notifications>

```

## Select a file datatransfer method

You can currently choose between FTP, HTTP, or HTTPS as the filetransfer method. The HTTP transfer method uses only a single copy per transfer, while the FTP method first copies the file to an FTP server and then copies it from there to the receiving side. Additionally, the HTTP transfer method reverts to simple filesystem copying whenever possible to optimize transfer speeds. However, to use HTTP transfers you must have ports open into most machines, which some may consider a security risk. The HTTPS transfer method meets this problem by having the HTTP communication secured and encrypted. To use the HTTPS transfer method you will need to generate a certificate that is needed to contact the embedded HTTPS server.

The FTP method requires one or more FTP-servers installed. (See Appendix A for further details). The XML below is a extract of a `settings.xml`, in which you have to replace `serverName`, `userName`, `userPassword` with proper values. This must be set for all applications.

```

<common>
  ...
  <remoteFile xsi:type="ftpremotefile">
    <!-- The class to use for RemoteFile objects. -->
    <class>dk.netarkivet.common.distribute.FTPRemoteFile</class>
    <!-- The default FTP-server used -->
    <serverName>hostname</serverName>
    <!-- The default FTP-server port used -->
    <serverPort>21</serverPort>
    <!-- The default FTP username -->
    <userName>exampleusername</userName>
    <!-- The default FTP password -->
    <userPassword>examplepassword</userPassword>
    <!-- The number of times FTPRemoteFile should try before giving up
        a copyTo operation. We augment FTP with checksum checks. -->
    <retries>3</retries>
  </remoteFile>

```



It is possible to use more than one FTP server, but each application can only use one. The FTP server that is used for a particular transfer is determined by the application that is sending a file. If you want to use more than one FTP-server, you must use different settings-files, or define the serverName and possibly also the userName and userPassword when starting the applications on the commandline with

```
-Dsettings.common.distribute.RemoteFile.serverName=FTP-server1
-Dsettings.common.distribute.RemoteFile.userName=ftpUser
-Dsettings.common.distribute.RemoteFile.userPassword=ftpPassword
```

Using HTTP as filetransfer method, you need to reserve a HTTP port on each machine per application. You can set this port on an application level on the commandline:

```
-Dsettings.common.remoteFile.port=5442
```

The following XML shows the the corresponding syntax in the settings.xml file:

```
<common>
  <remoteFile xsi:type="httpremotefile">
    <!-- The class to use for RemoteFile objects. -->
    <class>dk.netarkivet.common.distribute.HTTPRemoteFile</class>
    <!-- Port for embedded HTTP server -->
    <port>5442</port>
  </remoteFile>
```

Using the HTTPS file transfer method, you first need to generate a certificate that is used for communication. You can do this with the `keytool` application distributed with Java 5.

Run the following command:

```
keytool -alias NetarchiveSuite -keystore keystore -genkey
```

It should the respond with the following:

```
Enter keystore password:
```

Enter the password for the keystore.

The keytool will now prompt you for the following information

```
What is your first and last name?
[Unknown]:
What is the name of your organizational unit?
[Unknown]:
What is the name of your organization?
[Unknown]:
What is the name of your City or Locality?
[Unknown]:
What is the name of your State or Province?
[Unknown]:
What is the two-letter country code for this unit?
[Unknown]:
Is CN=Unknown, OU=Unknown, O=Unknown, L=Unknown, ST=Unknown, C=Unknown
correct?
[no]:
```

answer all the questions, and end with "yes".

Finally you will be asked for the certificate password.

```
Enter key password for <NetarchiveSuite>
(RETURN if same as keystore password):
```

Answer with a password for the certificate.

You now have a file called **keystore** which contains a certificate. This keystore needs to be available for all NetarchiveSuite applications, and referenced from settings as the following example shows:

```
<common>
...
<remoteFile xsi:type="httpsremotefile">
  <!-- The class to use for RemoteFile objects. -->
  <class>dk.netarkivet.common.distribute.HTTPSRemoteFile</class>
  <!-- The port for the remote file transfers -->
  <port>8300</port>
  <!-- The keystore -->
  <certificateKeyStore>path/to/keystore</certificateKeyStore>
  <!-- The keystore passwd -->
  <certificateKeyStorePassword>testpass</certificateKeyStorePassword>
  <!-- The key password-->
  <certificatePassword>testpass2</certificatePassword>
</remoteFile>
</common>
```

To keep your environment secure, you should make sure that the keystore and settings file *only* are readable for the user running the application.

## Configure job-generation

The scheduling takes place every one minute, unless the previous scheduling is not finished yet. The scheduling interval cannot be changed. Scheduling amounts to searching for active harvestdefinitions, that is ready to have jobs generated, and subsequently submitted for harvesting. The job-generation procedure are governed by a set of settings prefixed by *settings.harvester.scheduler..* These settings rule how large your crawljobs are going to be, and how long time they will take to complete. Note that harvestdefinitions consist of at least one DomainConfiguration, (containing a Heritrix setup, and a seed-list), and that there are two kinds: Snapshot Harvestdefinitions, and Selective Harvestdefinitions.

During scheduling, each harvest is split into a number of *crawl jobs*. This is done to keep Heritrix from using too much memory and to avoid that particularly slow or large domains cause harvests to take longer than necessary. In the job splitting part of the scheduling, the scheduler partitions a large number of DomainConfigurations into several crawljobs. Each crawljob can have only one Heritrix setup, so DomainConfigurations with different Heritrix setups will be split into different crawljobs. Additionally, a number of parameters influence what configurations are put into which jobs, attempting to create jobs that cover a reasonable amount of domains of similar sizes.

If you don't want to have the harvests split into multiple jobs, you just need to set each of `jobs.maxRelativeSizeDifference`, `jobs.minAbsoluteSizeDifference`, `jobs.maxTotalSize`, and `configChunkSize` to a large number, such as `MAX_LONG`. Initially, we suggest you don't change these parameters, as the way the work together is subtle. Harvests will always be split in different jobs, though, if they are based on different order.xml

templates, or if different harvest limits need to be enforced.

**errorFactorPrevResult:** Used when calculating expected size of a harvest of some domain during the job-creation process for snapshot harvests. This defines the factor by which we maximally allow domains that have previously been harvested to increase in size, compared to the value we estimate the domain to be. In other words, it defines how conservative our estimates are. The default value is 10, meaning that the maximum number of bytes harvested is as most 10 times as great as the value we use as expected size.

**errorFactorBestGuess:** Used when calculating expected size of a harvest of some domain during job-creation process for a snapshot Harvests. This defines the factor by which we maximally allow domains that have previously been incompletely harvested or not harvested at all to increase in size, compared to the value we estimate the domain to be. In other words, it defines how conservative our estimates are. The default value is 20, meaning that the maximum number of bytes harvested is as most 20 times as great as the value we use as expected size. This is probably an unreasonable number, it should be reset to 2 for most installations.

**expectedAverageBytesPerObject:** How many bytes the average object is expected to be on domains where we don't know any better. This number should grow over time, as of end of 2005 empirical data shows 38000. Default is 38000.

**maxDomainSize:** Initial guess of #objects in an unknown domain. Default value is 5000

**jobs.maxRelativeSizeDifference:** The maximum allowed relative difference in expected number of objects retrieved in a single job definition. Set to MAX\_LONG for no splitting.

**jobs.minAbsoluteSizeDifference:** Size differences for jobs below this threshold are ignored, regardless of the limits for the relative size difference. Set to MAX\_LONG for no splitting. Default value is 2000.

**jobs.maxTotalSize:** When this limit is exceeded no more configurations may be added to a job. Set to MAX\_LONG for no splitting. Default value is 2000000

**configChunkSize:** How many domain configurations we will process in one go before making jobs out of them. This amount of domains will be stored in memory at the same time. Set to MAX\_LONG for no job splitting. The default value is 10000.

MAX\_LONG refers to the number  $2^{63}-1$  or 9223372036854775807.

## Configure Domain Granularity

The NetarchiveSuite software is bound to the concept of Domains, where a Domain is defined as

```
"domainname"."tld"
```

This concept is useful for grouping harvests with regard to specific domains.

It can be configured what is considered a TLD by changing the settings files. The settings file currently distributed with the NetarchiveSuite software will list all country-level top-level-domains as "tld"s like ".dk", ".se" and ".no". However, as a proof of concept, for ".uk"-domains, there is listed the pseudo-top-level-domains ".co.uk", ".gov.uk", ".edu.uk" and some more.

Currently, only grouping by domain suffix is supported. A feature request is open for making the domain splitting pluggable. See [Feature Request 1072](#).

## Configure Heritrix process

Each harvester runs an instance of Heritrix for each harvest job being executed. It is possible to get access to the Heritrix web user interface for purposes of pausing or stopping a job, examining details of an ongoing harvest or even, if necessary, change an ongoing harvest. Note that some changes to harvests, especially those that change the scope and limits, may confuse the harvest definition system. We suggest using the Heritrix UI only for examination and pausing/terminating jobs.

Each harvest *application* running requires two ports, one for the user interface and one for JMX. The JMX port is set by the `settings.harvester.harvesting.heritrix.jmxPort` setting, and does not need to be open to other machines. The user interface port is set by the `settings.harvester.harvesting.heritrix.guiPort` setting, and should be open to the machines that the user interface should be accessible from. Make sure to have different ports for each harvest application if you're running more than one on a machine. Otherwise, your harvest jobs will fail when two harvest applications happen to try to run at the same time -- an error that could go unnoticed for a while, but which is more likely to happen exactly in critical situations where more harvesters are needed.

The Heritrix user interface is accessible through a browser using the port specified, e.g. <http://my.harvester.machine:8090>, and entering the administrator name and password set in the `settings.harvester.harvesting.heritrix.adminName` and `settings.harvester.harvesting.heritrix.adminPassword` settings. It is also possible to use JConsole to access the JMX interface of the Heritrix process, using one of the passwords in the `conf/jmxremote.password` file.

The final setting for the Heritrix processes is the amount of heap space each process is allowed to use. Since Heritrix uses a significant amount of heap space for seen URLs and other stuff, it is advisable to keep the `settings.harvester.harvesting.heritrix.heapSize` setting at at least its default setting of 1.5G if there is enough memory in the machine for this (remember to factor in the number of harvesters running on the machine -- swapping will slow the crawl down *significantly*).

## Configure web page look

The look of the web pages can be changed by changing files in the `webpages` directory. The files are distributed in war-files, which are simply zip-files. They can be unpacked to customize styles, and repacked afterwards using `zip`. Each of the five war files under `webpages` corresponds to one section of the web site, as seen in the left-hand menu. The two PNG files `transparent_logo.png` and `transparent_menu_logo.png` are used on the front page and atop the left-hand menu, respectively. They can be altered to suite your whim, but the width of `transparent_menu_logo.png` should not be increased so much that the menu becomes overly wide. The color scheme for each section is set in the `netarkivet.css` file for that section and can be changed to suit your whim, though we recommend changing them all at the same time to provide a uniform look.

## Configure security

Security in NetarchiveSuite is mainly defined in the `conf/security.policy` file. This file controls two main configurations: Which classes are allowed to do anything (core classes), and which classes are only allowed to read the files in the bit archive (third-party batch classes).

To enable the use of the security policy, you will need to launch your applications with the command line options `-Djava.security.manager` and `-Djava.security.policy=conf/security.policy`.

### Core classes

For the core classes, we need to identify all the classes that can be involved. The default `security.policy` file assumes that the program is started from the root of the distribution. If that is not the case, the `codeBase` entries must be changed to match. The following classes should be included:

- The `dk.netarkivet.*` jar files and supporting jar files, located in the `lib` directory. By default, all files in this directory and its subdirectories are included by the statement

```
grant codeBase "file:lib/-" {
    permission java.security.AllPermission;
};
```

- The `heritrix` jar files and supporting jar files for it, usually located in the `lib/heritrix/lib` directory. By default, these are included by the above, but in the QuickStart system, they are in a separate directory under `scripts/simple_harvest` (this place is included in the default `security.policy` file).
- The standard Java classes, which by default are included by the statement

```
grant codeBase "file:${java.home}/-" {
    permission java.security.AllPermission;
};
```

- The classes compiled by JSP as part of the web interface. These classes only exist on the machine(s) that run a web interface, and are found in the directory specified by the `settings.common.tempDir` setting. The default security file contains entries that assume this directory is `tests/commontempdir`. Note that an entry is required for each section of the web site:

```
grant codeBase "file:tests/commontempdir/Status/jsp/-" {
    permission java.security.AllPermission;
};
```

If you change the `settings.common.tempDir` setting, you will need to change this entry, too, or the web pages won't work.

### Third-party classes

The default `security.policy` file includes settings that allow third-party batch jobs to read the bitarchives set up for the QuickStart system. In a real installation, the bitarchive machines must specify which directories should be accessible and set up permissions for these. The default setup is:

```
grant {
  permission java.util.PropertyPermission
  "settings.archive.bitarchive.thisLocation", "read";
  permission java.io.FilePermission "${user.home}/netarchive/scripts
/simple_harvest/bitarchive1/filedir/*", "read";
  permission java.io.FilePermission "${user.home}/netarchive/scripts
/simple_harvest/bitarchive2/filedir/*", "read";
};
```

Notice how these permissions are not granted to a specific codebase, but the permissions given are very restrictive: The classes can read files in two explicitly stated directories, and can query for the value of the `settings.archive.bitarchive.thisLocation` setting -- all other settings are off-limits, as is reading and writing other files, including temporary files. If you wish to allow third-party batch jobs to do more, think twice first -- loopholes can be subtle.

## The actual deployment of the NetarchiveSuite

edit

This section describes one possible way to distribute the NetarchiveSuite software to a number of machines, but by no means the only one. In the examples below, we assume that `$NetarchiveSuiteDir` is set to the directory in which the NetarchiveSuite code is to be installed.

We assume that all machines in the chosen setup are unix/linux servers. The procedure below may not work on other platforms. After having created the new settings to be used in the deployment of the software, zip together the NetarchiveSuite files including the new settings and copy the modified NetarchiveSuite.zip to all machines taking part in the deployment:

```
export USER=test
export MACHINES="machine1.domain1, machine2.domain1, .. machine1.domain2,
machine2.domain2"
for MACHINE in $MACHINES; do
  scp NetarchiveSuite.zip $USER@$MACHINE:$NetarchiveSuiteDir
  ssh $USER@$MACHINE "cd $NetarchiveSuiteDir && unzip NetarchiveSuite.zip"
done
```

## Standard commandline settings

### The CLASSPATH

The CLASSPATH needed to start and run the java applications in NetarchiveSuite consists of 4 jarfiles, `dk.netarkivet.harvester.jar`, `dk.netarkivet.archive.jar`, `dk.netarkivet.viewerproxy.jar`, and `dk.netarkivet.monitor.jar`. The `dk.netarkivet.common.jar` and all our 3rd party dependencies need not be added explicitly to the CLASSPATH, as they are referenced indirectly in the jar-files.

```
export NetarchiveSuiteDir=/path/to/netarchiveSuite
export CLASSPATH=$CLASSPATH:$NetarchiveSuiteDir/lib
/dk.netarkivet.harvester.jar
export CLASSPATH=$CLASSPATH:$NetarchiveSuiteDir/lib
/dk.netarkivet.archive.jar
export CLASSPATH=$CLASSPATH:$NetarchiveSuiteDir/lib
/dk.netarkivet.viewerproxy.jar
export CLASSPATH=$CLASSPATH:$NetarchiveSuiteDir/lib
```

```
/dk.netarkivet.monitor.jar
```

## Logging

We use the `apache.commons.logging.framework`, so we need to point to the wanted logger-class (eg. `org.apache.commons.logging.impl.Jdk14Logger`) as well as to the logging configuration file. You may want to use different logging properties for different applications, especially when more than one application logs to the same logging directory. E.g. you want the change line `java.util.logging.FileHandler.pattern=./log/APPID%u.log` in the `conf/log.prop` file to something different.

```
export
LOG_SETTINGS="-Dorg.apache.commons.logging.Log=org.apache.commons.logging.
impl.Jdk14Logger \
-Djava.util.logging.config.file=$NetarchiveSuiteDir/conf/log.prop"
```

Note that if you use the `MonitorSiteSection`, your logging properties must contain the handler `dk.netarkivet.monitor.logging.CachingLogHandler`

## JMX settings

Each application has its own JMX- and RMI port. For instance the JMX port could be 8100 and the associated RMI port 8200, as in the example below, for the first machine on the host, then 8101/8201 for the second host, and so on. JMX also uses a password-file, which is the same throughout the installation (`$NetarchiveSuiteDir/conf/jmxremote.password`)

```
export JMX_SETTINGS="-Dsettings.common.jmx.port=8100 \
-Dsettings.common.jmx.rmiPort=8200"
```

## Select the appropriate settings.file for the application

The `conf/settings.xml` (the new one configured to your environment) is probably OK for most applications. But you may need to use special purpose settings-files for some applications, e.g. `BitarchiveApplications` (since you can't allocate more than one `fileDir` on the commandline). The settings file used in an application can be specified by:

```
export SETTING=-Ddk.netarkivet.settings.file=$NetarchiveSuiteDir/conf
/settings.xml
```

## JVM options

We need to set the maximum Java heap size to 1.5 Gbytes. You may use this to change that or add other JVM options.

```
export JAVA_OPTS=-Xmx1536m
```

## Admin machine

On the admin machine, we have to start the following 4 applications:

- 1 GUIApplication (Also controls the scheduler).
- 2 instances of BitarchiveMonitorApplication (Controlling the access to a bitarchive at a single location), one for each location (e.g. LocationA, and LocationB).
- 1 ARCRepositoryApplication (this application handles access to the bitarchives).

## Starting the GUIApplication

If you are not using the Embedded Derby Database, you can skip the following. This unzips the compressed fullhddb.jar located in \$NetarchiveSuiteDir/harvestdefinitionbasedir/:

```
cd $NetarchiveSuiteDir/harvestdefinitionbasedir
unzip fullhddb.jar
```

We also need to prepare the JSP-pages. You can unzip the war-files in the webpages directory as below:

```
cd $NetarchiveSuiteDir/webpages
rm -rf BitPreservation
unzip -o BitPreservation.war -d BitPreservation
rm -rf HarvestDefinition
unzip -o HarvestDefinition.war -d HarvestDefinition
rm -rf History
unzip -o History.war -d History
rm -rf QA
unzip -o QA.war -d QA
rm -rf Status
unzip -o Status.war -d Status
```

Or you can update your settings.xml file to refer to the war-files instead of the unpacked directories, for instance

```

        <siteSection>
            <!-- A subclass of SiteSection that defines this part of
the
                web interface. -->
<class>dk.netarkivet.harvester.webinterface.DefinitionsSiteSection</class>
            <!-- The directory or war-file containing the web
application
                for this site section.-->
<webapplication>webpages/HarvestDefinition.war</webapplication>
            <!-- The URL path for this section of the web interface.
-->
                <deployPath>/HarvestDefinition</deployPath>
        </siteSection>

```

and similar for other site sections.

Now we are ready to start the application:

```
cd $NetarchiveSuiteDir
export APP=dk.netarkivet.common.webinterface.GUIApplication
java $JAVA_OPTS $SETTING $LOG_SETTINGS $JMX_SETTINGS $APP
```

## Starting the BitarchiveMonitorApplication instances



In the general setup with a bitarchive distributed at two Locations, we have a BitarchiveMonitorApplication associated with each Location. Here the locations are "LocationA", and "LocationB". To distinguish the two instances from each other, we use the **settings.common.http.port** setting, which is used as a identifier (here we use 8081, 8082) as the two identifiers.

Start the monitor for bitarchive at 'LocationA' using "8081" as identifier thus:

```
cd $NetarchiveSuiteDir
export
APP_OPTIONS="-Dsettings.common.archive.bitarchive.thisLocation=LocationA \
-Dsettings.common.http.port=8081"
export APP=dk.netarkivet.archive.bitarchive.BitarchiveMonitorApplication
java $JAVA_OPTS $SETTING $LOG_SETTINGS $JMX_SETTINGS $APP_OPTIONS $APP
```

Start the monitor for the bitarchive at 'LocationB' using "8082" as identifier thus:

```
cd $NetarchiveSuiteDir
export
APP_OPTIONS="-Dsettings.common.archive.bitarchive.thisLocation=LocationB \
-Dsettings.common.http.port=8082"
export APP=dk.netarkivet.archive.bitarchive.BitarchiveMonitorApplication
java $JAVA_OPTS $SETTING $LOG_SETTINGS $JMX_SETTINGS $APP_OPTIONS $APP
```

- one ARCRepository (this application handles all access to the bitarchives).

```
cd $NetarchiveSuiteDir
export APP=dk.netarkivet.archive.arcrepository.ArcRepositoryApplication
java $JAVA_OPTS $SETTING $LOG_SETTINGS $JMX_SETTINGS $APP
```

## Harvester machines

On each harvester machine, we have one or more HarvestControllerApplications. Each HarvestControllerApplication have their own sidekick application that restarts the HarvestControllerApplication after each harvest. Settings related to the HarvestControllerApplication are

- settings.common.http.port (to distinguish between HarvestControllerApplications running on same machine)
- settings.harvester.harvesting.queuePriority (to select which of two queues to accept jobs from: HIGHPRIORITY (jobs part of a selective harvest), or LOWPRIORITY (jobs part of a snapshotharvest))
- settings.harvester.harvesting.minSpaceLeft (how many bytes *must* be available in the serverdir to accept crawljobs). The default is 400000000 (~400 Mbytes).

In the following, a low-priority HarvestControllerApplication is started with ID=8081

```
cd $NetarchiveSuiteDir
export
APP_OPTIONS="-Dsettings.harvester.harvesting.queuePriority=LOWPRIORITY \
-Dsettings.common.http.port=8081"
export
```

```
APP=dk.netarkivet.harvester.harvesting.HarvestControllerApplication
java $JAVA_OPTS $SETTING $LOG_SETTINGS $JMX_SETTINGS $APP_OPTIONS $APP
```

### Starting a SideKick application for the just started HarvestControllerApplication

```
cd $NetarchiveSuiteDir
export APP_OPTIONS=-Dsettings.common.http.port=8081
export APP=dk.netarkivet.harvester.sidekick.SideKick
export
APP_ARGS1=dk.netarkivet.harvester.sidekick.HarvestControllerServerMonitorH
ook
export APP_ARGS2=full or relative path to a script that can start the
HarvestControllerApplication again
java $JAVA_OPTS $SETTING $LOG_SETTINGS $JMX_SETTINGS $APP_OPTIONS $APP
$APP_ARGS1 $APP_ARGS2
```

## Bitarchive machines

For each Location, you can have BitarchiveServer's installed on one or more machines. We suggest using just one BitarchiveServer for each machine, though it is possible to use more than one. Each BitarchiveServer can have storage on several filesystems, so if archive-storage is spread over more than one filesystem, you need to modify the settings file like this

```
<settings>
..
<archive>
...
<bitarchive>
...
<fileDir>/home/bitarchiveOne/</fileDir>
<fileDir>/home/bitarchiveTwo/</fileDir>
...
</bitarchive>
</archive>
..
</settings>
```

Starting a BitarchiveServer requires knowing what Location it resides on, and the credentials required for correcting the data stored in the bitarchive:

```
cd $NetarchiveSuiteDir
export
APP_OPTIONS="-Dsettings.archive.bitarchive.thisLocation=ThisLocation \
-Dsettings.archive.bitarchive.thisCredentials=CREDENTIALS"
export APP=dk.netarkivet.archive.bitarchive.BitarchiveApplication
java $JAVA_OPTS $SETTING $LOG_SETTINGS $JMX_SETTINGS $APP_OPTIONS $APP
```

## Access servers

On the access-servers, we deploy any number of **ViewerProxyApplication** instances, and maybe one **IndexServerApplication** (only one in all) used to generate indices needed by the harvesters and the ViewerProxyApplication instances.

```
cd $NetarchiveSuiteDir
export APP=dk.netarkivet.archive.indexserver.IndexServerApplication
```

```
java $JAVA_OPTS $SETTING $LOG_SETTINGS $JMX_SETTINGS $APP
```

Each ViewerproxyApplication instance uses a dedicated HTTP port (settings.common.http.port), and its own distinct base directory (settings.viewerproxy.baseDir). They also belong to a Location (settings.archive.bitarchive.thisLocation). In the start sample below, the instance uses HTTP port 8081 and 'viewerproxy\_8081' as base directory, and belongs to locationA:

```
cd $NetarchiveSuiteDir
export APP_OPTIONS="-Dsettings.common.http.port=8081 \
-Dsettings.viewerproxy.baseDir=viewerproxy_8081 \
-Dsettings.archive.bitarchive.thisLocation=locationA"
export APP=dk.netarkivet.viewerproxy.ViewerProxyApplication
java $JAVA_OPTS $SETTING $LOG_SETTINGS $JMX_SETTINGS $APP_OPTIONS $APP
```

## Starting and stopping the NetarchiveSuite

edit

You need to start and stop the NetarchiveSuite applications in the correct order. The most critical part is that the BitarchiveMonitor must not start before the BitarchiveServers, as it might then try to perform batch jobs without all BitarchiveServers receiving the message. The following is a suggested order of startup:

### NetarchiveSuite application startup order

1. The BitarchiveApplication on all bitarchive servers is started:  
dk.netarkivet.archive.bitarchive.BitarchiveApplication
2. The applications on the admin-machine are started:

```
- dk.netarkivet.common.webinterface.GUIApplication
- dk.netarkivet.archive.arcrepository.ArcRepositoryApplication
- dk.netarkivet.archive.bitarchive.BitarchiveMonitorApplication for
Location A
- dk.netarkivet.archive.bitarchive.BitarchiveMonitorApplication for
Location B
```

3. The applications on the harvester machines are started: For each HarvesterControllerApplication deployed on this machine, the SideKick application is started, which in turn starts a HarvesterControllerApplication
4. The applications on the access-servers are started by first starting the IndexServer and then one or more ViewerproxyApplication instances.

### NetarchiveSuite application stopping order

After locating the process-id of any given process, the actually killing of the process is done on unix-machines with the command: `kill $PID`

The killing itself is done in the following order:

1. The applications on the admin-machine are killed:

```
- dk.netarkivet.common.webinterface.GUIApplication
- dk.netarkivet.archive.arcrepository.ArcRepositoryApplication
- dk.netarkivet.archive.bitarchive.BitarchiveMonitorApplication
```

2. The BitarchiveApplication on all bitarchive servers are shut down:

```
dk.netarkivet.archive.bitarchive.BitarchiveApplication
```

3. The applications on the harvester machines are shut down in this order: For each HarvesterControllerApplication deployed on this machine, the SideKick application is shut down first, and then the HarvesterControllerApplication. This prevents the SideKick application from starting a new HarvestControllerApplication.

4. The applications on the access-servers are shutdown by first killing the IndexServer and then the ViewerproxyApplication instances.

Remember to empty the JMS queues after shutting down the NetarchiveSuite if you are upgrading the system. If any outstanding JMS messages are around next time the NetarchiveSuite is started, they may cause deserialization errors if the message definitions have changed. To do this, you must recall what JMS environmentName your NetarchiveSuite instance have been using. The details of this are explained in Appendix A below.

In the Danish installation, we empty the queues each time the system is restarted, so the effect of leaving messages in the queues over a restart even when not upgrading has not been tested in practice.

## Monitoring an running instance of NetarchiveSuite

edit

The Status component of the NetarchiveSuite GUI implemented using JMX enables easy monitoring of all running applications. All important log messages (Log level INFO and above) can be studied in the GUI. However, only the last 100 messages from each application are available. This number can be increased or decreased using the setting *settings.monitor.logging.historySize*.

If you want to get more information about the current status of a particular application, you can use the program *jconsole*. You need to know on which machine the the application is running (HOSTNAME), the JMX port (JMX\_PORT) and RMI port (RMI\_PORT) assigned to the application, and password for the *monitorRole*. Then you just write *jconsole*, and click on the 'advanced' tab, enter the URL

```
service:jmx:rmi://HOSTNAME:RMI_PORT/jndi/rmi://HOSTNAME:JMX_PORT/jmxrmi
```

When asked for username, enter *monitorRole* and the password set for the application. Log machines can now be examined for the given application by seleting MBeans, and unfolding "dk.netarkivet.common.logging".

## Appendices

### Contents

1. Appendix A : Installing external software

1. Installing and configuring a JMS broker
  1. Obtaining JMS
  2. Installing the JMS broker
  3. Configuring the JMS broker
  4. Starting and stopping JMS
    1. How to empty queues
2. Installing and configuring FTP
  1. Starting and stopping a Proftpd server
2. Appendix B : Configurable settings in the NetarchiveSuite
3. Appendix C : Plugins in the NetarchiveSuite
4. Appendix D : Managing Heritrix harvest templates (order.xml)
  1. Mandatory elements in the NetarchiveSuite and their role
    1. 1) The QuotaEnforcer
    2. 2) The DeDuplicator
    3. 3) The "http-headers" element
    4. 4) The Archiver element
    5. 5) The Scope element
      1. The anatomy of a decidingscope
        1. A. The header
        2. B. The defining deciderule
        3. C. Standard harvest rules
        4. D. Define general crawlertraps to be avoided
  2. The HarvestTemplateApplication tool
  3. Predefined harvest templates
    1. Templates w/ DomainScope
    2. Templates w/ HostScope
    3. Templates w/ PathScope
5. Appendix E : Migrate the heritrix templates to NetarchiveSuite 3.6.0+
  1. How to convert from the former scopes to a decidingscope

## Appendix A : Installing external software

edit

### Installing and configuring a JMS broker

The software have been tested with the free JMS broker from Sun "Open Message Queue 4.1", and the commercial JMSBroker "Sun MQ 3.6 Enterprise Edition".

## Obtaining JMS

Sun's Open Message Queue can be obtained from the following site: <https://mq.dev.java.net/downloads.html> Go to the section named "Open Message Queue Binaries", and click on the Linux link in the subsection "Latest Open MQ 4.1 Binary Downloads". This will give you a jar-file named "mq4\_1-binary-Linux\_X86-XXXXXXXX.jar".

Note: We only support installation on the Linux platform here. However, you may want to install your JMS broker on a different platform. Binary versions are available at the site for: Solaris Sparc, Solaris x86, Linux (x86), Windows (x86). If you want to build a binary for another platform, the source can be downloaded from the download-page.

## Installing the JMS broker

Select Linux server where you want to install JMS broker, and select an installation directory. Then log on the linux server as root, and do the following:

```
export INSTALLATION_DIR=/path/to/installationdir
cd $INSTALLATION_DIR
unzip mq4_1-binary-Linux_X86-XXXXXXXX.jar
chmod +x ./mq/bin/imqbrokerd
./mq/bin/imqbrokerd -reset store -tty (tests that the broker can start
- CTRL-C to stop)
```

Check that it starts, and that the last message is "Broker <localhost>:7676 ready" We are now ready to configure the JMS broker.

## Configuring the JMS broker

- Edit the file `$INSTALLATION_DIR/mq/etc/imqenv.conf` to set `IMQ_DEFAULT_JAVAHOME` to a JDK1.5.0.
- Changing the number of the listening port number 7676 is done by editing the line `imq.portmapper.port=7676` in the file `$INSTALLATION_DIR/mq/lib/props/broker/default.properties`
- Set max listeners any given queue to 20. You need to make sure, that the following line `imq.autocreate.queue.maxNumActiveConsumers=20` is present and not commented out in the file `$INSTALLATION_DIR/mq/var/instances/imqbroker/props/config.properties`

(increase the number 20 if you have more than that number of applications of the same kind on the same location, for instance more than 20 bitarchiveapplications)

## Starting and stopping JMS

The broker is started directly in this way:

```
$INSTALLATION_DIR/mq/bin/imqbrokerd -reset store -tty &
```

The sysadmin would maybe like to start the broker on machine startup by inserting the statement above into the `/etc/rc.d/rc.local`

The broker is stopped in this way:

```
logon on machine as root
find processid for the broker (ps auxw | grep imqbrokerd)
kill -9 $IMQ_PROCESSID
```

Alternatively press Ctrl-c, if the terminal where the broker was started, is still available

You can test that JMS broker is alive by telnetting to its port, where it will give some technical information in reply:

```
[svc@udvikling kb-dev-adm-001.kb.dk]$ telnet localhost 7676
Trying 127.0.0.1...
Connected to localhost.localdomain (127.0.0.1).
Escape character is '^]'.
101 imqbroker 4.1
portmapper tcp PORTMAPPER 7676 [sessionid=1729683678303517696]
cluster_discovery tcp CLUSTER_DISCOVERY 46760
jmxrmi rmi JMX 0 [url=service:jmx:rmi://udvikling.kb.dk/stub/r00...Hg=]
admin tcp ADMIN 46763
jms tcp NORMAL 46762
cluster tcp CLUSTER 46764
.
Connection closed by foreign host.
```

To run JMS client applications, include the following jar files in the classpath :

`$INSTALLATION_DIR/mq/lib/jms.jar $INSTALLATION_DIR/mq/lib/imq.jar`

Create a passfile named `'imq_passfile'` (used when emptying JMS queues):

```
imq.imqcmd.password=REPLACE_WITH_PASSWORD
```

### How to empty queues

log on as root to the server, where the JMS broker is installed. The following assumes that the JMS environmentName is `PROD`, and that JMS password file resides in `~root/.imq_passfile`:

```
export JMS_ENV=PROD
export MQ_HOME=/usr/local
# imqcmd using -u admin -passfile ~/.imq_passfile
$MQ_HOME/bin/imqcmd list dst -t q -u admin -passfile ~/.imq_passfile |
grep ^${JMS_ENV}_ | cut -f1 -d\ |xargs -r -n 1 $MQ_HOME/bin/imqcmd
destroy dst -t q -u admin -passfile ~/.imq_passfile -f -n
$MQ_HOME/bin/imqcmd list dst -t t -u admin -passfile ~/.imq_passfile |
grep ^${JMS_ENV}_ | cut -f1 -d\ |xargs -r -n 1 $MQ_HOME/bin/imqcmd
destroy dst -t t -u admin -passfile
~/.imq_passfile -f -n"
```

## Installing and configuring FTP

If you decide to use FTPRemote for file transfer in the NetarchiveSuite, you need to install and start one or more FTP servers, before you begin the installation of the NetarchiveSuite. Any brand of FTP-servers will probably do, but we have good experience with Proftpd.

You can download Proftpd from <http://www.proftpd.org/>. We are using version 1.2.10, but any recent non-beta version will probably do.

The text below the proftpd.conf needed by NetarchiveSuite:

```
# Port 21 is the standard FTP port.
Port                21

# Umask 022 is a good standard umask to prevent new dirs and files
# from being group and world writable.
Umask               022

# To prevent DoS attacks, set the maximum number of child processes
# to 30.  If you need to allow more than 30 concurrent connections
# at once, simply increase this value.  Note that this ONLY works
# in standalone mode, in inetd mode you should use an inetd server
# that allows you to limit maximum number of processes per service
# (such as xinetd).
MaxInstances        30

# Set the user and group under which the server will run.
User                nobody
#Group              nogroup
Group               nobody

# To cause every FTP user to be "jailed" (chrooted) into their home
# directory, uncomment this line.
#DefaultRoot ~

# Normally, we want files to be overwriteable.
## This is necessary to allow the append operation
AllowOverwrite      on

AllowStoreRestart  on

# Bar use of SITE CHMOD by default
<Limit SITE_CHMOD>
  DenyAll
</Limit>

# This enables or disables the PAM authentication module.
# The default is 'on'.
#AuthPAM off
```

## Starting and stopping a Proftpd server

Log as root on to the server, where Proftpd is installed, and the following command will start the FTP-server

```
/usr/local/sbin/proftpd
```

The following will kill the FTP-server.

```
killall -9 proftpd
```



# Appendix B : Configurable settings in the NetarchiveSuite

edit

The NetarchiveSuite uses a XML file to define settings for its applications. The XML file must conform to the XML schema lib/data-definitions/settings.xsd. The XML file has a specific element for each of the dk.netarkivet.\* packages: common, harvester, archive, viewerproxy, and monitor.

In the common part of the settings.xml, we have general purpose settings (settings.common.tmpDir, settings.common.http.port), and settings, that allow us select plugins and their associated arguments(settings.common.RemoteFile.class, settings.common.jms.broker, settings.common.archiveRepositoryClient, and settings.common.indexClient.class).

In the harvester part of the settings.xml, we have settings configuring the harvesting process: scheduling, job splitting Most of these settings are used by the scheduler in DefinitionsSiteSection of the GUIApplication

In the archive part of the settings file, we have settings related to archive-access (e.g. certain timeouts, locations and their credentials is defined here). Also behaviour of the BitarchiveApplications is set here.

```
<?xml version="1.0" encoding="UTF-8"?>
<settings xmlns="http://www.netarkivet.dk/schemas/settings"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <common>
    <!-- Common temporary directory for all applications. Some
subdirs of
      this directory must be set to have AllPermission in the
conf/security.conf file, or the web pages won't work. -->
    <tempDir>./tests/commontempdir</tempDir>
    <!-- FTP connection data-->
    <remoteFile xsi:type="ftpremotefile">
      <!-- The class to use for RemoteFile objects. -->
      <class>dk.netarkivet.common.distribute.FTPRemoteFile</class>
      <!-- The default FTP-server used -->
      <serverName>localhost</serverName>
      <!-- The default FTP-server port used -->
      <serverPort>21</serverPort>
      <!-- The default FTP username -->
      <userName>exampleusername</userName>
      <!-- The default FTP password -->
      <userPassword>examplepassword</userPassword>
      <!-- The number of times FTPRemoteFile should try before
giving up
      a copyTo operation. We augment FTP with checksum checks.
-->
      <retries>3</retries>
    </remoteFile>
    <!-- Connection data for JMS-->
    <jms>
      <!-- Selects the broker class to be used. Must be a subclass
of
      dk.netarkivet.common.distribute.JMSConnection. -->
```

```

<class>dk.netarkivet.common.distribute.JMSConnectionSunMQ</class>
    <!-- The JMS broker host contacted by the JMS connection -->
    <broker>localhost</broker>
    <!-- The port the JMS connection should use -->
    <port>7676</port>
    <!-- The name of the environment in which this code is
running, e.g.
channels
        PROD, RELEASETEST, NHC,... Common prefix to all JMS
        -->
        <environmentName>DEV</environmentName>
</jms>
<http>
    <!-- The *unique* (per host) port number that may or may not
be
        used to serve http, but is frequently used to identify
        the process.-->
    <port>8076</port>
</http>
<arcrepositoryClient xsi:type="jmsarcrepositoryclient">
    <!-- The class that implements the ArcRepositoryClient. This
class
        will be instantiated by the ArcRepositoryClientFactory
-->
-->

<class>dk.netarkivet.archive.arcrepository.distribute.JMSArcRepositoryClie
nt</class>
    <!-- How many milliseconds we will wait before giving up on a
lookup request to the Arcrepository. Set to 1 minute to
make it possible to retrieve large records using FTP -->
    <getTimeout>60000</getTimeout>
    <!-- Number of times to try sending a store message before
failing,
        including the first attempt -->
    <storeRetries>3</storeRetries>
    <!-- Timeout in milliseconds before retrying when calling
ArcRepositoryClient.store() -->
    <storeTimeout>3600000</storeTimeout>
</arcrepositoryClient>
<monitorregistryClient xsi:type="jmsmonitorregistryclient">
    <!-- The class instantiated to register JMX urls at a
registry. -->
-->

<class>dk.netarkivet.common.distribute.monitorregistry.JMSMonitorRegistryC
lient</class>
    </monitorregistryClient>
    <indexClient xsi:type="indexrequestclient">
        <!-- The class instantiated to give access to indices. Will
be
            created by IndexClientFactory -->
-->

<class>dk.netarkivet.archive.indexserver.distribute.IndexRequestClient</cl
ass>
    <!-- The amount of time, in milliseconds, we should wait for
replies
        when issuing a call to generate an index over som jobs.
        -->
    <indexRequestTimeout>43200000</indexRequestTimeout>
</indexClient>
    <!-- The name of the directory where cache data global to the
entire
        machine can be stored. Various kinds of caches should be
stored in
        subdirectories of this -->
    <cacheDir>cache</cacheDir>
    <!--The number of milliseconds we wait for processes to react

```

```

        to shutdown requests.-->
<processTimeout>5000</processTimeout>
<!-- Error notification settings -->
<notifications>
    <!-- Which class to instantiate to handle error notifications
-->
        <class>dk.netarkivet.common.utils.EMailNotifications</class>
        <!-- The receiver of emails -->
        <receiver>example@netarkivet.dk</receiver>
        <!-- The stated sender of emails (and receiver of bounces)-->
        <sender>example@netarkivet.dk</sender>
</notifications>
<!-- Settings for sending email. Currently mail is only used for
email
    notifications. -->
<mail>
    <!-- The email server to use -->
    <server>examplesmtpserver.netarkivet.dk</server>
</mail>
<!-- JMX logging settings -->
<jmx>
    <!-- The port to connect to using JMX -->
    <port>8100</port>
    <!-- The RMI port used for communicating with beans -->
    <rmiPort>8200</rmiPort>
    <!-- The password file, containing information about who may
    connect -->
    <passwordFile>conf/jmxremote.password</passwordFile>
    <!-- How many seconds we will wait before giving up on a JMX
    connection. -->
    <timeout>120</timeout>
</jmx>
<!-- Settings for the web GUI -->
<webinterface>
    <!-- Language settings -->
    <language>
        <!-- A locale the GUI is available as -->
        <locale>da</locale>
        <!-- Native name of the language for this locale -->
        <name>Dansk</name>
    </language>
    <!-- Language settings -->
    <language>
        <!-- A locale the GUI is available as -->
        <locale>en</locale>
        <!-- Native name of the language for this locale -->
        <name>English</name>
    </language>
    <siteSection>
        <!-- A subclass of SiteSection that defines this part of
the
            web interface. -->
        <class>dk.netarkivet.harvester.webinterface.DefinitionsSiteSection</class>
        <!-- The directory or war-file containing the web
application
            for this site section.-->
        <webapplication>webpages/HarvestDefinition</webapplication>
        <!-- The URL path for this section of the web interface.
-->
        <deployPath>/HarvestDefinition</deployPath>
    </siteSection>
    <siteSection>
        <!-- A subclass of SiteSection that defines this part of
the

```

```

web interface. -->

<class>dk.netarkivet.harvester.webinterface.HistorySiteSection</class>
  <!-- The directory or war-file containing the web
application
      for this site section.-->
  <webapplication>webpages/History</webapplication>
  <!-- The URL path for this section of the web interface.
-->
  <deployPath>/History</deployPath>
</siteSection>
<siteSection>
  <!-- A subclass of SiteSection that defines this part of
the
      web interface. -->

<class>dk.netarkivet.archive.webinterface.BitPreservationSiteSection</clas
s>
  <!-- The directory or war-file containing the web
application
      for this site section.-->
  <webapplication>webpages/BitPreservation</webapplication>
  <!-- The URL path for this section of the web interface.
-->
  <deployPath>/BitPreservation</deployPath>
</siteSection>
<siteSection>
  <!-- A subclass of SiteSection that defines this part of
the
      web interface. -->

<class>dk.netarkivet.viewerproxy.webinterface.QASiteSection</class>
  <!-- The directory or war-file containing the web
application
      for this site section.-->
  <webapplication>webpages/QA</webapplication>
  <!-- The URL path for this section of the web interface.
-->
  <deployPath>/QA</deployPath>
</siteSection>
<siteSection>
  <!-- A subclass of SiteSection that defines this part of
the
      web interface. -->

<class>dk.netarkivet.monitor.webinterface.StatusSiteSection</class>
  <!-- The directory or war-file containing the web
application
      for this site section.-->
  <webapplication>webpages/Status</webapplication>
  <!-- The URL path for this section of the web interface.
-->
  <deployPath>/Status</deployPath>
</siteSection>
</webinterface>
</common>
<harvester>
  <datamodel>
    <domain>
      <!-- Default seed list to use when new domains are
created -->
      <defaultSeedlist>defaultseeds</defaultSeedlist>
      <!-- The name of a configuration that is created by
default and
          which is initially used for snapshot harvests-->
      <defaultConfig>defaultconfig</defaultConfig>

```

```
nothing      <!-- Name of order xml template used for domains if
this) -->      else is specified (e.g. newly created configurations use

<defaultOrderxml>default_orderxml</defaultOrderxml>
<!-- Default download rate for domain configuration.
Not currently enforced. -->
<defaultMaxrate>100</defaultMaxrate>
<!-- Default byte limit for domain configuration. -->
<defaultMaxbytes>1000000000</defaultMaxbytes>
<!-- This setting describes valid TLDs to define domains.
-->

<tld>ac</tld>
<tld>ad</tld>
<tld>ae</tld>
<tld>aero</tld>
<tld>af</tld>
<tld>ag</tld>
<tld>ai</tld>
<tld>al</tld>
<tld>am</tld>
<tld>an</tld>
<tld>ao</tld>
<tld>aq</tld>
<tld>ar</tld>
<tld>arpa</tld>
<tld>as</tld>
<tld>at</tld>
<tld>au</tld>
<tld>aw</tld>
<tld>ax</tld>
<tld>az</tld>
<tld>ba</tld>
<tld>bb</tld>
<tld>bd</tld>
<tld>be</tld>
<tld>bf</tld>
<tld>bg</tld>
<tld>bh</tld>
<tld>bi</tld>
<tld>biz</tld>
<tld>bj</tld>
<tld>bm</tld>
<tld>bn</tld>
<tld>bo</tld>
<tld>br</tld>
<tld>bs</tld>
<tld>bt</tld>
<tld>bv</tld>
<tld>bw</tld>
<tld>by</tld>
<tld>bz</tld>
<tld>ca</tld>
<tld>cat</tld>
<tld>cc</tld>
<tld>cd</tld>
<tld>cf</tld>
<tld>cg</tld>
<tld>ch</tld>
<tld>ci</tld>
<tld>ck</tld>
<tld>cl</tld>
<tld>cm</tld>
<tld>cn</tld>
<tld>co</tld>
<tld>com</tld>
```

<tld>coop</tld>  
<tld>cr</tld>  
<tld>cs</tld>  
<tld>cu</tld>  
<tld>cv</tld>  
<tld>cx</tld>  
<tld>cy</tld>  
<tld>cz</tld>  
<tld>de</tld>  
<tld>dj</tld>  
<tld>dk</tld>  
<tld>dm</tld>  
<tld>do</tld>  
<tld>dz</tld>  
<tld>ec</tld>  
<tld>edu</tld>  
<tld>ee</tld>  
<tld>eg</tld>  
<tld>eh</tld>  
<tld>er</tld>  
<tld>es</tld>  
<tld>et</tld>  
<tld>eu</tld>  
<tld>fi</tld>  
<tld>fj</tld>  
<tld>fk</tld>  
<tld>fm</tld>  
<tld>fo</tld>  
<tld>fr</tld>  
<tld>ga</tld>  
<tld>gb</tld>  
<tld>gd</tld>  
<tld>ge</tld>  
<tld>gf</tld>  
<tld>gg</tld>  
<tld>gh</tld>  
<tld>gi</tld>  
<tld>gl</tld>  
<tld>gm</tld>  
<tld>gn</tld>  
<tld>gov</tld>  
<tld>gp</tld>  
<tld>gq</tld>  
<tld>gr</tld>  
<tld>gs</tld>  
<tld>gt</tld>  
<tld>gu</tld>  
<tld>gw</tld>  
<tld>gy</tld>  
<tld>hk</tld>  
<tld>hm</tld>  
<tld>hn</tld>  
<tld>hr</tld>  
<tld>ht</tld>  
<tld>hu</tld>  
<tld>id</tld>  
<tld>ie</tld>  
<tld>il</tld>  
<tld>im</tld>  
<tld>in</tld>  
<tld>info</tld>  
<tld>int</tld>  
<tld>io</tld>  
<tld>iq</tld>  
<tld>ir</tld>  
<tld>is</tld>

<tld>it</tld>  
<tld>je</tld>  
<tld>jm</tld>  
<tld>jo</tld>  
<tld>jobs</tld>  
<tld>jp</tld>  
<tld>ke</tld>  
<tld>kg</tld>  
<tld>kh</tld>  
<tld>ki</tld>  
<tld>km</tld>  
<tld>kn</tld>  
<tld>kp</tld>  
<tld>kr</tld>  
<tld>kw</tld>  
<tld>ky</tld>  
<tld>kz</tld>  
<tld>la</tld>  
<tld>lb</tld>  
<tld>lc</tld>  
<tld>li</tld>  
<tld>lk</tld>  
<tld>lr</tld>  
<tld>ls</tld>  
<tld>lt</tld>  
<tld>lu</tld>  
<tld>lv</tld>  
<tld>ly</tld>  
<tld>ma</tld>  
<tld>mc</tld>  
<tld>md</tld>  
<tld>mg</tld>  
<tld>mh</tld>  
<tld>mil</tld>  
<tld>mk</tld>  
<tld>ml</tld>  
<tld>mm</tld>  
<tld>mn</tld>  
<tld>mo</tld>  
<tld>mobi</tld>  
<tld>mp</tld>  
<tld>mq</tld>  
<tld>mr</tld>  
<tld>ms</tld>  
<tld>mt</tld>  
<tld>mu</tld>  
<tld>museum</tld>  
<tld>mv</tld>  
<tld>mw</tld>  
<tld>mx</tld>  
<tld>my</tld>  
<tld>mz</tld>  
<tld>na</tld>  
<tld>name</tld>  
<tld>nc</tld>  
<tld>ne</tld>  
<tld>net</tld>  
<tld>nf</tld>  
<tld>ng</tld>  
<tld>ni</tld>  
<tld>nl</tld>  
<tld>no</tld>  
<tld>np</tld>  
<tld>nr</tld>  
<tld>nt</tld>  
<tld>nu</tld>

<tld>nz</tld>  
<tld>om</tld>  
<tld>org</tld>  
<tld>pa</tld>  
<tld>pe</tld>  
<tld>pf</tld>  
<tld>pg</tld>  
<tld>ph</tld>  
<tld>pk</tld>  
<tld>pl</tld>  
<tld>pm</tld>  
<tld>pn</tld>  
<tld>pr</tld>  
<tld>pro</tld>  
<tld>ps</tld>  
<tld>pt</tld>  
<tld>pw</tld>  
<tld>py</tld>  
<tld>qa</tld>  
<tld>re</tld>  
<tld>ro</tld>  
<tld>ru</tld>  
<tld>rw</tld>  
<tld>sa</tld>  
<tld>sb</tld>  
<tld>sc</tld>  
<tld>sd</tld>  
<tld>se</tld>  
<tld>sg</tld>  
<tld>sh</tld>  
<tld>si</tld>  
<tld>sj</tld>  
<tld>sk</tld>  
<tld>sl</tld>  
<tld>sm</tld>  
<tld>sn</tld>  
<tld>so</tld>  
<tld>sr</tld>  
<tld>st</tld>  
<tld>su</tld>  
<tld>sv</tld>  
<tld>sy</tld>  
<tld>sz</tld>  
<tld>tc</tld>  
<tld>td</tld>  
<tld>tf</tld>  
<tld>tg</tld>  
<tld>th</tld>  
<tld>tj</tld>  
<tld>tk</tld>  
<tld>tl</tld>  
<tld>tm</tld>  
<tld>tn</tld>  
<tld>to</tld>  
<tld>tp</tld>  
<tld>tr</tld>  
<tld>travel</tld>  
<tld>tt</tld>  
<tld>tv</tld>  
<tld>tw</tld>  
<tld>tz</tld>  
<tld>ua</tld>  
<tld>ug</tld>  
<tld>ac.uk</tld>  
<tld>co.uk</tld>  
<tld>gov.uk</tld>



```

<tld>ltd.uk</tld>
<tld>me.uk</tld>
<tld>mod.uk</tld>
<tld>net.uk</tld>
<tld>nic.uk</tld>
<tld>nhs.uk</tld>
<tld>org.uk</tld>
<tld>plc.uk</tld>
<tld>police.uk</tld>
<tld>sch.uk</tld>
<tld>govt.uk</tld>
<tld>orgn.uk</tld>
<tld>lea.uk</tld>
<tld>mil.uk</tld>
<tld>nel.uk</tld>
<tld>uk</tld>
<tld>us</tld>
<tld>uy</tld>
<tld>uz</tld>
<tld>va</tld>
<tld>vc</tld>
<tld>ve</tld>
<tld>vg</tld>
<tld>vi</tld>
<tld>vn</tld>
<tld>vu</tld>
<tld>wf</tld>
<tld>ws</tld>
<tld>ye</tld>
<tld>yt</tld>
<tld>yu</tld>
<tld>za</tld>
<tld>zm</tld>
<tld>zw</tld>
</domain>
<database xsi:type="derbydatabase">
  <!-- The full URL for connecting to the database.
        If present and not empty, this URL must match the
settings
        baseDir and class.-->
  <url>jdbc:derby:harvestdefinitionbasedir/fullhddb</url>
  <!-- The class that handles DB-specific methods -->

<specificsclass>dk.netarkivet.harvester.datamodel.DerbyEmbeddedSpecifics</
specificsclass>
0..24
  <!-- The earliest time of day backup will be initiated,
backup
        hours. At a time shortly after this, a consistent
        copy of the database will be created -->
  <backupInitHour>3</backupInitHour>
</database>
</datamodel>
<scheduler>
  <!-- Used when calculating expected size of a harvest of some
how
        configuration during job-creation process. This defines
larger
        great a possible factor we will permit a harvest to be
previous
        then the expectation, when basing the expectation on a
        completed job. -->
  <errorFactorPrevResult>10</errorFactorPrevResult>
  <!-- Used when calculating expected size of a harvest of some
how
        configuration during job-creation process. This defines

```

```

larger          great a possible factor we will permit a harvest to be
previous       then the expectation, when basing the expectation on
               uncompleted harvests or no harvest data at all. -->
               <errorFactorBestGuess>20</errorFactorBestGuess>
domains        <!-- How many bytes the average object is expected to be on
over           where we don't know any better. This number should grow
               over
               time, as of end of 2005 empirical data shows 38000 -->
<expectedAverageBytesPerObject>38000</expectedAverageBytesPerObject>
               <!-- Initial guess of #objects in an unknown domain -->
               <maxDomainSize>5000</maxDomainSize>
               <jobs><!-- One Job corresponds to a Heritrix run -->
               <!-- The maximum allowed relative difference in expected
number         of objects retrieved in a single job definition.
Set to        MAX_LONG for no splitting -->
               <maxRelativeSizeDifference>100</maxRelativeSizeDifference>
               <!-- Size differences for jobs below this threshold are
ignored,      regardless of the limits for the relative size
difference.   Set to MAX_LONG for no splitting. -->
               <minAbsoluteSizeDifference>2000</minAbsoluteSizeDifference>
               <!-- When this limit is exceeded no more configurations
may be       added to a job. Set to MAX_LONG for no splitting. -->
               <maxTotalSize>2000000</maxTotalSize>
               </jobs>
before        <!-- How many domain configurations we will process in one go
stored       making jobs out of them. This amount of domains will be
               in memory at the same time. Set to MAX_LONG for no job
               splitting. -->
               <configChunkSize>10000</configChunkSize>
               </scheduler>
               <harvesting>
and           <!-- Each job gets a subdir of this dir. Job data is written
               Heritrix writes to that subdir-->
               <serverDir>server</serverDir>
               <!-- The minimum amount of free bytes in the serverDir
               required before accepting any harvest-jobs. Default is
               400000000 bytes (~400 Mbytes).
               -->
               <minSpaceLeft>400000000</minSpaceLeft>
               <!-- The directory in which data from old jobs is kept after
to           uploading. Each directory from serverDir will be moved
               here if any data remains, either due to failed uploads or
               because it wasn't attempted uploaded. -->
               <oldjobsDir>oldjobs</oldjobsDir>
               <!-- Pool to take jobs from -->
               <queuePriority>HIGHPRIORITY</queuePriority>
               <!-- When to stop Heritrix, timeouts in ms. -->
               <heritrix>
               <!-- The timeout setting for aborting a crawl based on
this         crawler-inactivity. If the crawler is inactive for
               amount of seconds the crawl will be aborted.

```

```

        The inactivity is measured on the
        crawlController.activeToeCount(). -->
        <inactivityTimeout>100</inactivityTimeout>
        <!-- The timeout value (in seconds) used in
HeritrixLauncher
        for aborting crawl when no bytes are being received
from
        web servers. -->
        <noresponseTimeout>100</noresponseTimeout>
        <!-- Name for accessing the Heritrix GUI -->
        <adminName>admin</adminName>
        <!-- Password for accessing the Heritrix GUI -->
        <adminPassword>adminPassword</adminPassword>
        <!-- Port used to access the Heritrix web user interface.
machine.
        This port must not be used by anything else on the
        machine.
        -->
        <guiPort>8090</guiPort>
        <!-- Port used to access the Heritrix JMX interface.
machine,
        This port must not be used by anything else on the
machines
        but does not need to be accessible from other
        unless you want to be able to use jconsole to access
        Heritrix directly
        -->
        <jmxPort>8091</jmxPort>
        <!-- The heap size to use for the Heritrix sub-process.
This
        should probably be fairly large. It can be
specified in
        the same way as for the -Xmx argument to Java, e.g.
        512M, 2G etc.-->
        <heapSize>1598M</heapSize>
        </heritrix>
        <!-- The file used to signal that the harvest controller is
running.
        Sidekick starts HarvestController if this file is not
present
        -->
        <isrunningFile>./hcsRunning.tmp</isrunningFile>
        </harvesting>
        </harvester>
        <archive>
        <arcrepository>
        <!-- Absolute/relative path to where the "central list of
files and
        checksums" (admin.data) is written. Used by
ArcRepository and
        BitPreservation. -->
        <baseDir>.</baseDir>
        <!-- The names of all bit archive locations in the
environment, e.g., "KB" and "SB". -->
        <location>
        <name>SB</name>
        </location>
        <location>
        <name>KB</name>
        </location>
        <!-- Default bit archive to use for batch jobs (if none is
specified) -->
        <batchLocation>KB</batchLocation>
        </arcrepository>
        <bitarchive>
        <!-- The minimum amount of bytes left *in any dir* that we
will

```

```

allow a bitarchive machine to accept uploads with. When
no
dir has more space than this, the bitarchive machine
stops
listening for uploads. This values should at the very
least
be greater than the largest ARC file you expect to
receive.
-->
<minSpaceLeft>200000000</minSpaceLeft>
<!-- These are the directories where ARC files are stored
from
(in a subdir). If more than one is given, they are used
one end. -->
<fileDir>m:\bitarchive</fileDir>
<fileDir>n:\bitarchive</fileDir>
<fileDir>o:\bitarchive</fileDir>
<fileDir>p:\bitarchive</fileDir>
<!-- The frequency in milliseconds of heartbeats that are
sent by
each BitarchiveServer to the BitarchiveMonitor. -->
<heartbeatFrequency>1000</heartbeatFrequency>
<!-- If we haven't heard from a bit archive within this many
wait
milliseconds, we don't expect it to be online and won't
for them to reply on a batch job. This number should be
for
significantly greater than heartbeatFrequency to account
temporary network congestion. -->
<acceptableHeartbeatDelay>60000</acceptableHeartbeatDelay>
<!-- The BitarchiveMonitorServer will listen for
BatchEndedMessages
for this many milliseconds before it decides that a
batch job
is taking too long and returns just the replies it has
received at that point. -->
<batchMessageTimeout>1209600000</batchMessageTimeout>
<!-- For archiving applications, which bit archive are you
part of?-->
<thisLocation>SB</thisLocation>
<!-- Credentials to enter in the GUI for "deleting" ARC files
in
this bit archive -->
<thisCredentials>examplecredentials</thisCredentials>
<!-- When the length record exceeds this number, the contents
of the record
will be transferred using a RemoteFile. Currently set to
10 MB
-->
<limitForRecordDatatransferInFile>10485760</limitForRecordDatatransferInFi
le>
</bitarchive>
<bitpreservation>
<!-- Absolute or relative path to dir containing results of
file-list-batch-jobs and checksumming batch jobs
for bit preservation-->
<baseDir>bitpreservation</baseDir>
</bitpreservation>
</archive>
<viewerproxy>
<!-- The main directory for the ViewerProxy, used for storing the
Lucene
index for the jobs being viewed -->
<baseDir>viewerproxy</baseDir>
</viewerproxy>

```

```

<monitor>
  <!-- The name of the application, fx.
"BitarchiveServerApplication".
  The monitor puts this with each log message -->
  <applicationName>NA</applicationName>
  <logging>
    <historySize>100</historySize>
  </logging>
</monitor>
</settings>

```

## Appendix C : Plugins in the NetarchiveSuite

edit All the settings above ending on ".class" indicate that the implementation of a certain feature can be replaced by alternative implementations. There is usually a choice of several classes to choose from, but this is not always the case. But our framework does at least enable the installer to replace the default class with a class of his own, if no existing alternative suits.

We now describe the available plugs, and existing plugins for these plugs.

**settings.common.remoteFile.class:** This setting allows you to select your chosen way of filetransfer in the NetarchiveSuite. You can here choose between FTPRemoteFile (where the data is transferred using a FTP-server), HTTPRemoteFile (where the data is transferred using a two embedded webservers (one at each end), and HTTPSRemoteFile which works just like HTTPRemoteFile except it uses a shared certificate file for secure communication. **Firewall-note** The HTTPRemoteFile and HTTPSRemoteFile requires dedicated ports to be open between all possible senders and recipients of data. For implementers of new filetransfer methods, this class must implement the class `dk.netarkivet.common.distribute.RemoteFile`. The default value is FTPRemoteFile.

**settings.harvester.datamodel.database.specificsclass:** This setting allows you select which type of database you want to use. There are support for 3 types already: An Embedded Derby database (`dk.netarkivet.harvester.datamodel.DerbyEmbeddedSpecifics`), an external Derby database (`dk.netarkivet.harvester.datamodel.DerbyClientSpecifics`), or an MySQL database (`dk.netarkivet.harvester.datamodel.MySQLSpecifics`). The default is `DerbyEmbeddedSpecifics`. If you choose not to use the default, you need to replace the default database URL(setting `settings.harvester.datamodel.database.url`), and maybe the time for the daily backup to start (setting `settings.harvester.datamodel.database.backupInitHour`)

**settings.common.jms.class** This class designates what kind of JMS broker the NetarchiveSuite uses to send messages between applications. Presently only the Sun JMS brokers is supported (`dk.netarkivet.common.distribute.JMSConnectionSunMQ`). This class must implement the `dk.netarkivet.common.distribute.JMSConnection` class.

**settings.common.arcrepositoryClient.class.** Must implement `dk.netarkivet.common.distribute.ArcRepositoryClient` The available choices are the default `dk.netarkivet.archive.arcrepository.distribute.JMSArcRepositoryClient` (that is required, if you want to access the distributed type of archive that is included in the NetarchiveSuite). and the `dk.netarkivet.common.distribute.TrivialArcRepositoryClient` (allows for access to a local archive)

**settings.common.notifications.class:** Allows for different ways of making notifications. The default choice is the class `dk.netarkivet.common.utils.EmailNotifications`

(which allows you to receive notifications by email). The use of this plugin requires setting the mail-server, the recipient- and sending email-address. Alternatively, you can use `dk.netarkivet.common.utils.PrintNotifications`, which simply prints the notifications to `stderr` on the terminal.

**settings.common.webinterface.sitesection.class** This setting allows you to add webmodules to the NetarchiveSuite GUI. Several SiteSection classes can be active in the same GUI. the default(standard) configuration contains all 5 existing webmodules:

1. HarvestDefinition: Allows you to define and schedule harvests ,
2. HarvestHistory: See the status of running and finished harvestjobs
3. BitPreservation: This module has tools for sanity testing data in the bitarchives
4. QA: Module for doing Quality Assurance
5. Status: Module for monitoring the health of all machines and applications

**settings.common.webinterface.language:** The languages supported by the webinterface. Danish (locale=da)and English (locale=en) are supported currently. The Developer Manual will tell you how to add support for more languages to the NetarchiveSuite.

**settings.common.indexClient:** The client selected for access to indices. Indices are requested by the HarvesterControllerApplication instances.

```
<indexClient xsi:type="indexrequestclient">
  <!-- The class instantiated to give access to indices. Will
be
          created by IndexClientFactory -->

<class>dk.netarkivet.archive.indexserver.distribute.IndexRequestClient</cl
ass>
  <!-- The amount of time, in milliseconds, we should wait for
replies
          when issuing a call to generate an index over som jobs.
-->
  <indexRequestTimeout>43200000</indexRequestTimeout>
</indexClient>
```

## Appendix D : Managing Heritrix harvest templates (order.xml)

edit

The NetarchiveSuite software uses a patched version of Heritrix 1.12.1 to harvest webpages. A harvest done by Heritrix is specified with a harvest template (invariably named order.xml). A harvest template describes how much to harvest and from where. Furthermore a seedlist is always associated with a given order.xml.

The standard harvest template used by NetarchiveSuite follow the order.xml standard of Heritrix 1.10+.

Our default harvest template can be seen here in full: [📄 default\\_orderxml.xml](#)

If you intend to build your own templates, it is recommended to use this template as a baseline.

## Mandatory elements in the NetarchiveSuite and their role

A number of elements in the order.xml are required in all NetarchiveSuite harvest templates:

### 1) The QuotaEnforcer

The QuotaEnforcer is used to restrict the number of bytes harvested from each domain in the seedlist.

```
<newObject name="QuotaEnforcer"
class="org.archive.crawler.prefetch.QuotaEnforcer">
  <boolean name="force-retire">false</boolean>
  <boolean name="enabled">true</boolean>
  <newObject name="QuotaEnforcer#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
    <map name="rules">
      </map>
  </newObject>
  <long name="server-max-fetch-successes">-1</long>
  <long name="server-max-success-kb">-1</long>
  <long name="server-max-fetch-responses">-1</long>
  <long name="server-max-all-kb">-1</long>
  <long name="host-max-fetch-successes">-1</long>
  <long name="host-max-success-kb">-1</long>
  <long name="host-max-fetch-responses">-1</long>
  <long name="host-max-all-kb">-1</long>
  <long name="group-max-fetch-successes">-1</long>
  <long name="group-max-success-kb">-1</long>
  <long name="group-max-fetch-responses">-1</long>
  <long name="group-max-all-kb">-1</long>
</newObject>
```

### 2) The DeDuplicator

The DeDuplicator is a module authored by Kristinn Sigurdsson from the National Library of Iceland. It is part of the [Heritrix](#) Write-processor chain. It enables us to avoid saving duplicates in our storage. It does this by looking up the url of the potential duplicate object in the index associated with this module. If the url is found in the index, and the checksum for the url in the index is unaltered, the object is not stored. However a reference to where the object is stored is written to the crawl log. If the url for the object is not found in the index, the object is stored normally. Note that only non-text objects are examined by this module, i.e. where the mimetype of the object does not match `^text/.*` (like text/html or text/plain).

NetarchiveSuite uses a patched version of the 0.3.0-20061218 beta version of the deduplicator.


```
<newObject name="DeDuplicator"
class="is.hi.bok.deduplicator.DeDuplicator">
  <boolean name="enabled">true</boolean>
  <map name="filters">
  </map>
  <string name="index-location"/>
  <string name="matching-method">By URL</string>
  <boolean name="try-equivalent">true</boolean>
  <boolean name="change-content-size">>false</boolean>
```

```

<string name="mime-filter">^text/.*</string>
<string name="filter-mode">Blacklist</string>
<string name="analysis-mode">Timestamp</string>
<string name="log-level">SEVERE</string>
<string name="origin"/>
<string name="origin-handling">Use index information</string>
<boolean name="stats-per-host">true</boolean>
</newObject>

```

### 3) The "http-headers" element

This element describes, how Heritrix will present itself to the webservers when fetching data. It points by default to the non-existing webpage [http://my\\_website.com/my\\_infopage.html](http://my_website.com/my_infopage.html) and the equally non-existing mail address "  my\_email@my\_website.com ". Please update this to your own institution and email!

```

<map name="http-headers">
  <string name="user-agent">Mozilla/5.0 (compatible;
heritrix/1.12.1 +http://my_website.com/my_infopage.html)</string>
  <string name="from">my_email@my_website.com</string>
</map>

```

### 4) The Archiver element

This element does the actual writing of the fetched objects to an arcfile. In the future we may want to write to WARC files instead, which can be easily be done. Heritrix allows you to have multiple 'Writers' in use at the same time. For instance, you can write your objects to both ARC and WARC at the same time, as well as writing the objects to a database.

```

<newObject name="Archiver"
class="org.archive.crawler.writer.ARCWriterProcessor">
  <boolean name="enabled">true</boolean>
  <newObject name="Archiver#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
    <map name="rules">
    </map>
  </newObject>
  <boolean name="compress">>false</boolean>
  <string name="prefix">IAH</string>
  <string name="suffix">${HOSTNAME}</string>
  <integer name="max-size-bytes">100000000</integer>
  <stringList name="path">
    <string>arcs</string>
  </stringList>
  <integer name="pool-max-active">5</integer>
  <integer name="pool-max-wait">300000</integer>
  <long name="total-bytes-to-write">0</long>
  <boolean name="skip-identical-digests">>false</boolean>
</newObject>

```

### 5) The Scope element

The scope element decides which urls to harvest and which not to harvest. Before release 3.6.0, we used the following three scopes:

- A. DomainScope. The standard NetarchiveSuite scope allows the harvester to fetch all objects coming from any 2nd level domains represented by one of the seeds. Embedded objects,



like images, and stylesheets are always fetched even when coming from other domains.

B. HostScope. This scope are restricted to fetching objects from the hosts represented by the seeds.

C. PathScope. This scope are restricted to fetching objects from

These 3 scopes were all deprecated from Heritrix 1.10.0, and now all NetarchiveSuite templates are required to use the DecidingScope instead. This type of Scope uses a sequence of DecideRules to define the scope of the harvest. We now emulate these three scopes by adding a specific DecideRule to the DecidingScope. In the case of DomainScope, it required designing our own DecideRule (dk.netarkivet.harvester.harvesting.OnNSDDomainsDecideRule). So for DomainScope type scopes, you add the following element:

```
<newObject name="acceptURIFromSeedDomains"  
class="dk.netarkivet.harvester.harvesting.OnNSDDomainsDecideRule">  
    <string name="decision">ACCEPT</string>  
    <string name="surts-source-  
file">seeds.txt</string>  
    <boolean name="seeds-as-surt-  
prefixes">>false</boolean>  
    <string name="surts-dump-file"/>  
    <boolean name="also-check-  
via">>false</boolean>  
    <boolean name="rebuild-  
on-reconfig">>true</boolean>  
</newObject>
```

Emulating the HostScope requires adding the OnHostsDecideRule element:

```
<newObject name="acceptIfOnSeedsHosts"  
class="org.archive.crawler.deciderules.OnHostsDecideRule">  
    <string name="decision">ACCEPT</string>  
    <string name="surts-dump-file"></string>  
    <boolean name="also-check-  
via">>false</boolean>  
    <boolean name="rebuild-  
on-reconfig">>true</boolean>  
</newObject>
```

Emulating the PathScope requires adding the SurtPrefixesDecideRule element:

```
<newObject name="acceptIfSurtPrefixed"  
class="org.archive.crawler.deciderules.SurtPrefixedDecideRule">  
    <string name="decision">ACCEPT</string>  
    <string name="surts-source-file"></string>  
    <boolean name="seeds-as-surt-  
prefixes">>true</boolean>  
    <string name="surts-dump-file"></string>  
    <boolean name="also-check-  
via">>false</boolean>  
    <boolean name="rebuild-  
on-reconfig">>true</boolean>  
</newObject>
```

An example of a complete DecidingScope element is shown below.

```
<newObject name="scope"  
class="org.archive.crawler.deciderules.DecidingScope">
```

```

<boolean name="enabled">true</boolean>
<string name="seedsfile">seeds.txt</string>
<boolean name="reread-seeds-on-config">true</boolean>
<!-- DecideRuleSequence. Multiple DecideRules applied in
order with last non-PASS the resulting decision -->
<newObject name="decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
  <map name="rules">
    <newObject name="rejectByDefault"
class="org.archive.crawler.deciderules.RejectDecideRule"/>
    <newObject name="acceptURIFromSeedDomains"
class="dk.netarkivet.harvester.harvesting.OnNSDomainsDecideRule">
      <string name="decision">ACCEPT</string>
      <string name="surts-source-file"></string>
      <boolean name="seeds-as-surt-
prefixes">true</boolean>
      <string name="surts-dump-file"/>
      <boolean name="also-check-
via">false</boolean>
      <boolean name="rebuild-
on-reconfig">true</boolean>
    </newObject>
    <newObject name="rejectIfTooManyHops"
class="org.archive.crawler.deciderules.TooManyHopsDecideRule">
      <integer name="max-hops">25</integer>
    </newObject>
    <newObject name="rejectIfPathological"
class="org.archive.crawler.deciderules.PathologicalPathDecideRule">
      <integer name="max-
repetitions">3</integer>
    </newObject>
    <newObject name="acceptIfTranscluded"
class="org.archive.crawler.deciderules.TransclusionDecideRule">
      <integer name="max-trans-
hops">25</integer>
      <integer name="max-speculative-
hops">1</integer>
    </newObject>
    <newObject name="pathdepthfilter"
class="org.archive.crawler.deciderules.TooManyPathSegmentsDecideRule">
      <integer name="max-
path-depth">20</integer>
    </newObject>
    <newObject name="global_crawlertraps"
class="org.archive.crawler.deciderules.MatchesListRegExpDecideRule">
      <string name="decision">REJECT</string>
      <string name="list-logic">OR</string>
      <stringList name="regexp-list">
        <string>.*core\.UserAdmin.*core\.UserLogin.*
</string>
        <string>.*core\.UserAdmin.*register
\.UserSelfRegistration.*</string>
        <string>.*\w\/index
\.php\?title=Speci[ae]l:Recentchanges.*</string>
        <string>.*act=calendar&cal_id=.*</string>
        <string>.*advCalendar_pi.*</string>
        <string>.*cal\.asp\?date=.*</string>
        <string>.*cal\.asp\?view=monthly&date=.*
</string>
        <string>.*cal\.asp\?view=weekly&date=.*
</string>
        <string>.*cal\.asp\?view=yearly&date=.*
</string>
        .....
        <string>.*index\.php\?iDate=.*</string>
        <string>.*index\.php\?module=PostCalendar&

```

```

amp;func=view.*</string>
                                <string>.*index\.php\?option=com_events&
amp;task=view.*</string>
                                <string>.*index\.php\?option=com_events&
amp;task=view_day&year=.*</string>
                                <string>.*index\.php\?option=com_events&
amp;task=view_detail&year=.*</string>
                                <string>.*index\.php\?option=com_events&
amp;task=view_month&year=.*</string>
                                <string>.*index\.php\?option=com_events&
amp;task=view_week&year=.*</string>
                                </stringList>
                                </newObject>
                                </map> <!-- end rules -->
                                </newObject> <!-- end decide-rules -->
                                </newObject> <!-- End DecidingScope -->

```

## The anatomy of a decidingscope

Finally, we describe the rest of the components of a decidingscope element.

### A. The header

```

<newObject name="scope"
class="org.archive.crawler.deciderules.DecidingScope">
    <boolean name="enabled">true</boolean>
    <string name="seedsfile">seeds.txt</string>
    <boolean name="reread-seeds-on-config">true</boolean>
    <!-- DecideRuleSequence. Multiple DecideRules applied in
order with last non-PASS the resulting decision -->
    <newObject name="decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
        <map name="rules">
            <newObject name="rejectByDefault"
class="org.archive.crawler.deciderules.RejectDecideRule"/>

```

### B. The defining deciderule

Here we have the deciderule, that defines this as either a DomainScope, a HostScope, or a PathScope

### C. Standard harvest rules

These rules add more restrictions to the scope:

- Restrict the amount of hops allowed from any seed. Normally set to 25.
- Restrict the amount of repetitions in a URL-path, eg. repetition/repetition/... Repetitions are normally symptoms of crawlertraps.
- Define the maximal transclusion hops, and maximal speculative hops. (🌐 <http://crawler.archive.org/apidocs/org/archive/crawler/deciderules/TransclusionDecideRule.html>)
- Restrict the maximal path depth. Normally set to 20

```

                                <newObject name="rejectIfTooManyHops"
class="org.archive.crawler.deciderules.TooManyHopsDecideRule">

```

```

                <integer name="max-hops">25</integer>
            </newObject>
            <newObject name="rejectIfPathological"
class="org.archive.crawler.deciderules.PathologicalPathDecideRule">
                <integer name="max-
repetitions">3</integer>
            </newObject>
            <newObject name="acceptIfTranscluded"
class="org.archive.crawler.deciderules.TransclusionDecideRule">
                <integer name="max-trans-
hops">25</integer>
                <integer name="max-speculative-
hops">1</integer>
            </newObject>
            <newObject name="pathdepthfilter"
class="org.archive.crawler.deciderules.TooManyPathSegmentsDecideRule">
                <integer name="max-
path-depth">20</integer>
            </newObject>

```

#### D. Define general crawlertraps to be avoided

Lists of crawlertraps to be avoided are defined with a `MatchesListRegExpDecideRule`. Here we list all crawlertraps (defined by a regular expression). If any object matches one of these regular expression, the object is not fetched (unless a previous rule requires the object to be fetched).

```

<newObject name="global_crawlertraps"
class="org.archive.crawler.deciderules.MatchesListRegExpDecideRule">
    <string name="decision">REJECT</string>
    <string name="list-logic">OR</string>
    <stringList name="regexp-list">
        <string>.*core\.UserAdmin.*core
\.UserLogin.*
    </stringList>

```

When creating a new `Harvestjob`, another `MatchesListRegExpDecideRule` is added to the `harvestTemplate`, that specifies the crawlertraps to be avoided.

## The HarvestTemplateApplication tool

You can upload and download the templates using our GUI. This is described in our [User Manual](#). But you can also upload and download the templates using the commandline `HarvestTemplateApplication`. This application allows you to create, download, update templates. We have made a script to make it easier to use this application: [📄](#)  
`HarvestTemplateApplication.sh.txt`

```

java dk.netarkivet.harvester.datamodel.HarvestTemplateApplication
<command> <args>
create <template-name> <xml-file for this template>
download [<template-name>]
update <template-name> <xml-file to replace this template>
showall

```

## Predefined harvest templates

All our templates fall in three categories depending on the scope defined in the template. Note that

our templates generally do not obey robots.txt. This is because the Danish legislation allows is to ignore the constraints dictated by robots.txt. However, there are two exceptions to this rule:

- default\_obeyrobots.xml
- default\_obeyrobots\_withforms.xml

Even though DomainScope, HostScope, PathScope are now emulated using DecidingScope, these categories are still useful:

### **Templates w/ DomainScope**

1. default\_orderxml.xml (standard template)
2. default\_withforms.xml (standard template that can handle forms)
3. default\_obeyrobots.xml (standard template that can handle forms)
4. default\_obeyrobots\_withforms.xml (standard template that obeys robots.txt and handles forms)
5. default\_orderxml\_low\_bandwidth.xml (standard template for sites with low bandwidth)
6. frontpages.xml (harvest template that only harvest the seeds and associated stylesheets and images)
7. frontpages\_plus\_1level.xml (The above plus one extra level extra)
8. frontpages\_plus\_2levels.xml (The above plus 2 extra levels)

### **Templates w/ HostScope**

1. host\_10levels\_orderxml.xml (harvest the hosts of the seeds up to 10 levels from seeds)
2. host\_100levels\_orderxml.xml (harvest the hosts of the seeds up to 100 levels from seeds)

### **Templates w/ PathScope**

1. path\_10levels\_orderxml.xml (harvest the hosts of the seeds up to 10 levels from seeds)
2. path\_100levels\_orderxml.xml (harvest the hosts of the seeds up to 100 levels from seeds)

## **Appendix E : Migrate the heritrix templates to NetarchiveSuite 3.6.0+**

If you are just using the predefined templates with few changes like changed the email-address and website information, the easiest way to migrate is to modify the predefined templates found in the binary distribution of NetarchiveSuite in the harvestdefinitionbasedir/order\_templates\_dist directory and change the email-adress and website information again.

If you do this, you also get the more inconsequential updates to the template:

- The removal of obsolete attributes from some elements
- Addition of new attributes to some elements

Then you just update the existing templates in your database with these modified ones using the HarvestTemplateApplication tool mentioned in AppendixD. Note that some templates are no longer distributed with NetarchiveSuite. If you want to keep using those, you need to follow the procedure described below.

If you have already put a lot effort in making your own templates, you can update your existing templates by "only" upgrading the scope element in the templates from either a DomainScope, HostScope, or a PathScope.

Before we explain how to migrate these scopes to a DecidingScope, you need to know something about the anatomy of these scopes.

1) Header (includes scope class, and attributes):

```
<newObject name="scope" class="org.archive.crawler.scope.PathScope">
  <boolean name="enabled">true</boolean>
  <string name="seedsfile">seeds.txt</string>
  <boolean name="reread-seeds-on-config">true</boolean>
  <integer name="max-link-hops">10</integer>
  <integer name="max-trans-hops">5</integer>
```

2) An OrFilter element named "exclude-filter" containing a number of filters as components: a HopsFilter, a PathDepthFilter, a PathologicalPathFilter, a URIRegExpFilter, a URIListRegExpFilter (filter to avoid common crawlertraps), and potentially other types of filters: Each of these filters will have to be converted to a similar DecideRule. Explanation to follow.

```
      <newObject name="exclude-filter"
class="org.archive.crawler.filter.OrFilter">
  <boolean name="enabled">true</boolean>
  <boolean name="if-matches-return">true</boolean>
  <map name="filters">
    <newObject name="hops_filter"
class="org.archive.crawler.filter.HopsFilter">
      <boolean name="enabled">true</boolean>
    </newObject>
    <newObject name="pathdepth"
class="org.archive.crawler.filter.PathDepthFilter">
      <boolean name="enabled">true</boolean>
      <integer name="max-path-depth">20</integer>
      <boolean name="path-less-or-equal-
return">false</boolean>
    </newObject>
    <newObject name="pathologicalpath"
class="org.archive.crawler.filter.PathologicalPathFilter">
      <boolean name="enabled">true</boolean>
      <integer name="repetitions">3</integer>
    </newObject>
    <newObject name="dr_dk"
class="org.archive.crawler.filter.URIRegExpFilter">
      <boolean name="enabled">true</boolean>
      <boolean name="if-match-return">true</boolean>
      <string name="regexp">.*dr\.dk.*epg\.asp.*
</string>
    </newObject>
    <newObject name="globale_crawlertraps"
class="org.archive.crawler.filter.URIListRegExpFilter">
      <boolean name="enabled">true</boolean>
      <boolean name="if-match-return">true</boolean>
      <string name="list-logic">OR</string>
      <stringList name="regexp-list">
        <string>.*core\.UserAdmin.*core\.UserLogin.*
</string>
        <string>.*core\.UserAdmin.*register
\.UserSelfRegistration.*</string>
        <string>.*\w\/index
\.php\?title=Speci[ae]l:Recentchanges.*</string>
```

```

        <string>.*act=calendar&cal_id=.*</string>
        .....
        <string>.*calendar\.asp\?qMonth=.*</string>
        <string>.*calendar\.php\?sid=.*</string>
        <string>.*worldscinet\.com.*</string>
        <string>.*www3\.interscience\.wiley\.com.*
</string>
        <string>.*www-gdz\.sub\.uni-goettingen\.de.*
</string>
        </stringList>
    </newObject>
</map>
</newObject>

```

3) Additional filters. Here we have a "Force-accept-filter", an "additionalScopeFocus" filter, and a "transitive Filter", of which only the transitiveFilter element needs to be converted. The two other elements are just deleted.

```

    <newObject name="force-accept-filter"
class="org.archive.crawler.filter.OrFilter">
    <boolean name="enabled">true</boolean>
    <boolean name="if-matches-return">true</boolean>
    <map name="filters">
    </map>
</newObject>
    <newObject name="additionalScopeFocus"
class="org.archive.crawler.filter.FilePatternFilter">
    <boolean name="enabled">true</boolean>
    <boolean name="if-match-return">true</boolean>
    <string name="use-default-patterns">All</string>
    <string name="regexp"/>
</newObject>
    <newObject name="transitiveFilter"
class="org.archive.crawler.filter.TransclusionFilter">
    <boolean name="enabled">true</boolean>
    <integer name="max-speculative-hops">1</integer>
    <integer name="max-referral-hops">15</integer>
    <integer name="max-embed-hops">15</integer>
</newObject>
</newObject> <!-- end of scope element -->

```

## How to convert from the former scopes to a decidingscope

Converting the header is easy. All headers have the form:

```

<newObject name="scope"
class="org.archive.crawler.deciderules.DecidingScope">
    <boolean name="enabled">true</boolean>
    <string name="seedsfile">seeds.txt</string>
    <boolean name="reread-seeds-on-config">true</boolean>
    <!-- DecideRuleSequence. Multiple DecideRules applied in
order with last non-PASS the resulting decision -->
    <newObject name="decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
        <map name="rules">
            <newObject name="rejectByDefault"
class="org.archive.crawler.deciderules.RejectDecideRule"/>

```

plus a special defining deciderule that emulates the DomainScope, the HostScope, or the PathScope. 1) The defining deciderule for DomainScope is (the only one using a special purpose DecideRule):

```
<newObject name="acceptURIFromSeedDomains"  
class="dk.netarkivet.harvester.harvesting.OnNSDomainsDecideRule">  
    <string name="decision">ACCEPT</string>  
    <string name="surts-source-  
file">seeds.txt</string>  
    <boolean name="seeds-as-surt-  
prefixes">>false</boolean>  
    <string name="surts-dump-file"/>  
    <boolean name="also-check-  
via">>false</boolean>  
    <boolean name="rebuild-  
on-reconfig">>true</boolean>  
</newObject>
```

2) The defining deciderule for HostScope is:

```
<newObject name="OnHostsRule"  
class="org.archive.crawler.deciderules.OnHostsDecideRule">  
    <string name="decision">ACCEPT</string>  
    <string name="surts-dump-file"/>  
    <boolean name="also-check-via">>false</boolean>  
    <boolean name="rebuild-on-reconfig">>true</boolean>  
</newObject>
```

3) The defining deciderule for PathScope is:

```
<newObject name="acceptIfSurtPrefixed"  
class="org.archive.crawler.deciderules.SurtPrefixedDecideRule">  
    <string name="decision">ACCEPT</string>  
    <string name="surts-source-file"></string>  
    <boolean name="seeds-as-surt-  
prefixes">>true</boolean>  
    <string name="surts-dump-file"></string>  
    <boolean name="also-check-  
via">>false</boolean>  
    <boolean name="rebuild-  
on-reconfig">>true</boolean>  
</newObject>
```

After the header and the defining deciderule, we add a deciderule corresponding to the 'hops\_filter'. Note that the two last attributes 'max-link-hops', and 'max-trans-hops' in the header cease to be general scope attributes. Instead max-trans-hops become an attribute for the "acceptIfTranscluded" mentioned above, and the 'max-link-hops' attribute becomes an attribute for the new 'hops\_filter' deciderule. The following

```
<integer name="max-link-hops">10</integer>  
<newObject name="hops_filter"  
class="org.archive.crawler.filter.HopsFilter">  
    <boolean name="enabled">>true</boolean>  
</newObject>
```



is then translated to the following deciderule

```
<newObject name="rejectIfTooManyHops"  
class="org.archive.crawler.deciderules.TooManyHopsDecideRule">  
  <integer name="max-hops">10</integer>  
</newObject>
```

Following this, we need to add a translation of the 'pathdepth' element, and the 'pathologicalpath' element, plus a translation of the 'transitiveFilter' element in the last part of the scope. The following

```
  <newObject name="pathdepth"  
class="org.archive.crawler.filter.PathDepthFilter">  
    <boolean name="enabled">true</boolean>  
    <integer name="max-path-depth">20</integer>  
    <boolean name="path-less-or-equal-  
return">false</boolean>  
  </newObject>  
  <newObject name="pathologicalpath"  
class="org.archive.crawler.filter.PathologicalPathFilter">  
    <boolean name="enabled">true</boolean>  
    <integer name="repetitions">3</integer>  
  </newObject>  
  
  <newObject name="transitiveFilter"  
class="org.archive.crawler.filter.TransclusionFilter">  
    <boolean name="enabled">true</boolean>  
    <integer name="max-speculative-hops">1</integer>  
    <integer name="max-referral-hops">15</integer>  
    <integer name="max-embed-hops">15</integer>  
  </newObject>
```

is translated to

```
<newObject name="rejectIfPathological"  
class="org.archive.crawler.deciderules.PathologicalPathDecideRule">  
  <integer name="max-repetitions">3</integer>  
</newObject>  
<newObject name="acceptIfTranscluded"  
class="org.archive.crawler.deciderules.TransclusionDecideRule">  
  <integer name="max-trans-hops">5</integer>  
  <integer name="max-speculative-hops">1</integer>  
</newObject>  
<newObject name="pathdepthfilter"  
class="org.archive.crawler.deciderules.TooManyPathSegmentsDecideRule">  
  <integer name="max-path-depth">20</integer>  
</newObject>
```

Note that the attributes 'max-referral-hops' and 'max-embed-hops' in the 'transitiveFilter' element have been merged into one single attribute 'max-trans-hops' which is now no longer an attribute of the scope, as it was in the old scopes.

Now you only need to convert all remaining URIRegExpFilter and URIListRegExpFilter

elements to a corresponding DecideRule. The deciderule corresponding to URIRegExpFilter is MatchesRegExpDecideRule, and the deciderule corresponding to URIListRegExpFilter is MatchesListRegExpDecideRule. Converting the dr\_dk element (a URIRegExpFilter)

```
<newObject name="dr_dk"
class="org.archive.crawler.filter.URIRegExpFilter">
    <boolean name="enabled">true</boolean>
    <boolean name="if-match-return">true</boolean>
    <string name="regexp">.*dr\.dk.*epg\.asp.*
</string>
</newObject>
```

gives us:

```
<newObject name="dr_dk"
class="org.archive.crawler.deciderules.MatchesRegExpDecideRule">
    <string name="decision">REJECT</string>
    <string name="regexp">.*dr\.dk.*epg\.asp.*</string>
</newObject>
```

Converting the globale\_crawlertraps element (URIListRegExpFilter)

```
<newObject name="globale_crawlertraps"
class="org.archive.crawler.filter.URIListRegExpFilter">
    <boolean name="enabled">true</boolean>
    <boolean name="if-match-return">true</boolean>
    <string name="list-logic">OR</string>
    <stringList name="regexp-list">
        <string>.*core\.UserAdmin.*core\.UserLogin.*
</string>
        <string>.*core\.UserAdmin.*register
\.UserSelfRegistration.*</string>
        <string>.*\w\/index
\.php\?title=Speci[ae]l:Recentchanges.*</string>
        <string>.*act=calendar&cal_id=.*</string>
        .....
        <string>.*calendar\.asp\?qMonth=.*</string>
        <string>.*calendar\.php\?sid=.*</string>
        <string>.*worldscinet\.com.*</string>
        <string>.*www3\.interscience\.wiley\.com.*
</string>
        <string>.*www-gdz\.sub\.uni-goettingen\.de.*
</string>
    </stringList>
</newObject>
```

gives us

```
<newObject name="globale_crawlertraps"
class="org.archive.crawler.deciderules.MatchesListRegExpDecideRule">
    <string name="decision">REJECT</string>
    <string name="list-logic">OR</string>
    <stringList name="regexp-list">
        <string>.*core\.UserAdmin.*core\.UserLogin.*</string>
        <string>.*core\.UserAdmin.*register
\.UserSelfRegistration.*</string>
        <string>.*\w\/index
\.php\?title=Speci[ae]l:Recentchanges.*</string>
        <string>.*act=calendar&cal_id=.*</string>
    </stringList>
</newObject>
```

```
.....
    <string>.*calendar\.asp\?qMonth=.*</string>
    <string>.*calendar\.php\?sid=.*</string>
    <string>.*worldscinet\.com.*</string>
    <string>.*www3\.interscience\.wiley\.com.*</string>
  </stringList>
</newObject>
```

Finally we need to wrap up the the sequence of deciderules and the scope itself. So we add

```
    </map> <!-- end rules -->
  </newObject> <!-- end decide-rules -->
</newObject> <!-- End DecidingScope -->
```