

The NWA Toolset Manual

\$Id: nwaToolsetManual.xml,v 1.4 2003/12/17 12:50:52 sverreb Exp \$

The NWA Toolset Manual

Copyright © 2003 by Royal Library in StockholmRoyal Library in CopenhagenHelsinki University Library in FinlandNational Library of NorwayNational and University Library of Iceland

Table of Contents

About this Guide	i
Purpose of this guide.....	i
Audience	i
1. Installation	1
Requirements	1
Installer Requirements.....	1
Retriever Requirements	2
Exporter Requirements.....	2
Search Engine Requirements.....	2
Access Module Requirements	2
Installation.....	2
Retriever Installation	3
Exporter Installation	4
Access Module Installation	5
Search Engine Installation.....	6
2. Starting to use the NWA Toolset	7
Testing the Retriever	7
Compiling an aid-list for the Exporter	7
Exporting.....	8
Indexing.....	8
Searching.....	8
3. Configuring	9
Exporter configuration	9
Access Module configuration.....	9

List of Examples

1-1. Retriever install dialogue.....	4
1-2. Exporter install dialogue	4
1-3. Access Module install dialogue.....	5
1-4. Exporter install dialogue	6

About this Guide

Purpose of this guide

The purpose of this guide is to describe and aid the installation process of the NWA Toolset.

Audience

This guide's intended audience are those who actually will carry out the installation of the NWA Toolset.

Chapter 1. Installation

In order to install the NWA Toolset you should download the latest NWA Toolset distribution. You may obtain the NWA Toolset from this location: <http://nwa.nb.no/>

The different components of the Toolset may require installation and/or configuration of additional software packages. See the next chapters for details on this.

The installation steps are shown below.

- Install the Retriever. The Retriever is the interface between a Web Archive and the NWA Toolset components Exporter and Access Module. In order to export the contents of a Web Archive to the NWA Document Format (to feed the search engines indexer), the presence of a working Retriever is an absolute Requirement. If you don't need to interface your own archive right away but just want to try out the Toolset, you may install the default Retriever and our tiny sample Web Archive along with it.
- Install the Exporter. The Exporter is the Toolset component that fetches the archived files from the Web Archive (through the Retriever), extracts data from them and stores these data in the NWA Document Format, ready for indexing
- Install the Access Module. The Access Module is the Toolset component that will enable you to search the indexed archive data and navigate the Web Archive.
- Install the Search Engine. The Search Engine is not a Toolset component as such. The NWA Toolset has been designed so that it should be possible to interface different search engines. Our specification for the NWA search engine abstraction layer enables anyone to implement an interface to their search engine of choice without having to modify the NWA Toolset. The NWA Toolset distribution does however include the Apache Jakarta Lucene search engine adapted for NWA Toolset use. Install this, along with our tiny sample Web Archive if you want to try out the NWA Toolset without implementing anything of your own.

Note: The different NWA Toolset components does not have to be installed on the same machine. See the chapter on Configuration (yet to come, sorry) for information on how the components communicates with each other.

Requirements

The NWA Toolset and the NWA Adapted Lucene search engine has been tested on **Linux RedHat 7.3**. No attempt has been made to verify whether the NWA Toolset may work on other Operating System versions.

Installer Requirements

In order to run the install script a working version of **Perl** has to be installed.

Information about installing Perl may be found at <http://www.perl.org>

Note: Please note that even if the installer may be invoked using any version of Perl, the Exporter requires a specific Perl version.

Retriever Requirements

The Retriever for Nedlib style web archives requires an **Apache HTTP server** (v.1.3 or later) with PHP (v.4.3.1 or later)

Information about installing Apache with PHP may be found at <http://httpd.apache.org/> and <http://www.php.net/>

Exporter Requirements

The Exporter requires Perl v.5.8.2 or later is installed including the following CPAN modules

- XML::TokeParser
- HTML::Parser (When installing this module accept the default “no” on the question “do you want decoding on unicode entities”)
- HTML::TokeParser
- LWP::Simple
- URI
- HTTP::date
- Text::Iconv
- Getopt::Std
- POSIX

Information about installing Perl and the CPAN modules may be found at <http://www.perl.org> and <http://www.cpan.org>

Search Engine Requirements

The NWA Toolset adapted Lucene search engine requires the following:

- JDK 1.4 (or later).
- Apache Jakarta Tomcat 4.1.18 (or later). You can find information of how to install Apache Jakarta Tomcat server at <http://java.sun.com> and <http://jakarta.apache.org/>.

You can find information of how to install JDK and Apache Jakarta Tomcat server at <http://java.sun.com>, <http://java.sun.com> and <http://jakarta.apache.org/>.

Access Module Requirements

The Access Module requires that the Apache HTTP server (v 1.3 or later) with mod-perl (perl 5.8 or later) and PHP (v 4.3.1 or later). The apache configuration directive 'AllowOverride' in httpd.conf must be set to at least 'Options FileInfo'.

Information about installing Apache with PHP and Perl may be found at <http://httpd.apache.org/>, <http://www.php.net/> and <http://www.perl.org>

Installation

In order to install the NWA Toolset follow the instructions given below:

- Download the `nwatoolset_<buildno>.tar.gz` from the NWA web site to the directory where you want to run the installation from.
- If you intend to install the sample archive as well download `nwa_samplearchive.tar.gz` from <http://nwa.nb.no> to the host where you want to put the Retriever and unpack it using `tar xvfz nwa_samplearchive.tar.gz` in a directory of your choice.
- Unpack the NWA Toolset using `tar xvfz nwatoolset_<buildno>.tar`
- Invoke the installation dialogue by typing `./install_nwatoolset.pl`. You may have to alter the path-name immediately following the sha-bang (#!) at the head of `nwa_toolset.pl` script or invoke the script using `perl install_nwatoolset.pl`
- Follow the instructions given by the install dialogue. Below is shown the output generated when invoking the install script.

```
# ./install_nwatoolset.pl
```

```
NWA TOOLSET INSTALLATION
```

```
This installation script will install the NWA Toolset or
selected NWA Toolset components.
```

```
If you want to distribute the different components on different
machines choose the components to install on this machine and re-execute
the install script on the other machine(s) where you want
the other component(s).
```

```
Note that the NWA Toolset requires a search engine as well.
See the Installation Guide for information on how to install
the NWA adapted Lucene Search Engine or see System Configuration Guide
on how to interface your own search engine of choice.
```

```
Type in numbers of components you want to install:
```

1. Retriever
2. Exporter
3. Access module
4. Lucene search engine

You may choose to install all or some of the Toolset components. Simply type in the numbers corresponding to the components you want to install separated by space. The following sections gives detail information about the installation steps.

If you want to distribute the different Toolset components to different hosts, then you should repeat the installation process on every host in question.

Retriever Installation

The install script will ask you to enter the following information:

- Where to put the Retriever script. This has to be a directory that is part of the HTTP servers web tree.

Below is shown an example of the Retriever install dialogue.

Example 1-1. Retriever install dialogue

```
STARTING INSTALLATION OF NEDLIB RETRIEVER
```

```
For each question during installation, hit Enter to accept  
a default value presented inside [], or type in a new value.
```

```
Nedlib Retriever should be installed into a directory  
where you can execute php-scripts. Type in the name of a suitable  
php-directory and Retriever will be placed into php-directory/nwatooset  
Default for php-dir is [/var/www/html]:
```

```
Creating directory /var/www/html/nwatooset
```

```
Nedlib Retriever has been installed into /var/www/html/nwatooset.
```

```
Finished installation of Nedlib retriever
```

Exporter Installation

The Exporter installation will ask you to submit the following information.

- The directory in which to install the Exporter. Make sure that the user invoking the script has sufficient permissions to this directory, as well as other directories asked for later in the install process.
- The Retriever URL.
- The directory where you want the output from the Exporter stored
- A name for your collection. This value will be “stamped” on every record exported giving you the possibility to distinguish different parts of the archive from others when later querying the index. In the cooperation between the Nordic National Libraries this field so far has been used for country code (no, se, fi, dk and is)

Below is shown an example of the Exporter install dialogue.

Example 1-2. Exporter install dialogue

```
STARTING INSTALLATION OF EXPORTER
```

```
For each question during installation, hit Enter to accept a default  
value presented inside [], or type in a new value.
```

```
Type in installation directory for Exporter.  
This directory will by default contain subdirectories for  
source code, configuration and log files  
[default /usr/local/nwatooset]:
```

```
Creating directory /usr/local/nwatooset
```

```
Creating directory /usr/local/nwatoolset/bin
Creating directory /usr/local/nwatoolset/conf
Creating directory /usr/local/nwatoolset/log
```

```
Exporter gets documents from Web archive through Document Retriever.
Type in URL for the Retriever, usually it should be something like
http://localhost/nwatoolset/docretriever_nedlib.php
if you have installed Retriever into an nwatoolset subdirectory
under php directory on your machine.
Default for Retriever URL is [http://test.nb.no/nwatoolset/docretriever_nedlib.php]:
```

```
Type in the absolute name of the nwa directory into which
Exporter will write nwa file(s)
default is [/data/nwa_dir]:/home/test/data
```

```
Type in the name of your collection
default is [test]:
```

```
Finished installation of Exporter files. They were placed into
/home/test/bin and /home/test/conf
```

```
The exporters log file will be placed in the directory /usr/local/nwatoolset/log/
```

Access Module Installation

The install script will ask you to enter the following information:

- Install path. Where to install the Access Module. This has to be a directory in the web servers web tree.
- What search engine to use. Choose Lucene if you don't have a FAST Search Engine ready for use.
- The url to the Document Retriever
- The url of this installation (depends on your web server configuration as well as the directory you chose to install the Access Module into).
- The collection name used in the index. Set this to the same as the collection name set in the Exporter installation.
- The Lucene Search Engine URL

Below is shown an example of the Access Module install dialogue.

Example 1-3. Access Module install dialogue

```
STARTING INSTALLATION OF ACCESS MODULE
```

```
For each question during installation, hit Enter to accept
a default value presented inside [], or type in a new value.
```

```
Type in rootpath for Access module. The Access module will be
installed into 'rootpath/nwatoolset'
default for rootpath is [/var/www/html]:
```

```
Type in the URL to Document Retriever,
default is [http://test.nb.no/nwatoolset/docretriever_nedlib.php]:
```

Type in URL for this installation,
default is [http://test.nb.no/nwatoolset]:

Type in the name of your collection (same as used in Exporter installation)
default is [test]:

You are using Lucene search engine. Type in URL for
luceneconf_searchengineurl, or leave blank for default
default is [http://test.nb.no:8080/nwa/servlet/nwa]:

Installed Access module into /var/www/html/nwatoolset

Search Engine Installation

The install dialogue for the NWA adapted Lucene search engine will ask you for the absolute path to your TomCat servers webapps directory. See install dialogue example below.

Example 1-4. Exporter install dialogue

STARTING INSTALLATION OF LUCENE SEARCH ENGINE

Type in the full path of the TomCat installation directory
or leave blank if you have no TC yet: /usr/local/jakarta-tomcat/

Unpacked the Nwa Lucene Web application to /usr/local/jakarta-tomcat/webapps

Chapter 2. Starting to use the NWA Toolset

In order to start using the NWA Toolset you need to perform the steps outlined below.

1. Test that the Retriever is functioning correctly
2. Compile a list of documents to export
3. Perform an export using the list above as input
4. Index the output from the Exporter

These steps are described in more detail below.

Testing the Retriever

In order to test the Retriever try accessing the following urls in a browser (or use **wget [URL]** from the command line):

- `http://<hostname>.<domainname>[:port]/<retrieverpath>/docretriver_nedlib.php?reqtype=getmeta&aid=<archivepath>12-01/1/ebec0183244bb27d77989efc5be91218`
- `http://<hostname>.<domainname>[:port]/<retrieverpath>/docretriver_nedlib.php?reqtype=getfile&aid=<archivepath>12-01/1/ebec0183244bb27d77989efc5be91218`

Where *retrieverpath* is the path to where the Retriever was installed (relative to web tree root) and *archivepath* is the absolute path to where the sample Web Archive were stored

The first HTTP request (getmeta) should return archived technical metadata for the file with the given archive identifier (aid). An example of such metadata is given below.

```
<?xml version="1.0" encoding="UTF-8" ?>
<metadata>
  <url><![CDATA[http://www.nb.no/test/1.html]]></url>
  <time>20031201082345</time>
  <contenttype>
    <type>text/html</type>
    <charset>ISO-8859-1</charset>
  </contenttype>
  <http-header><![CDATA[
    HTTP/1.1 200 OK
    Date: Fri, 1 Dec 2003 00:02:56 GMT
    Server: Apache/1.3.26 (Unix) PHP/4.1.2
    X-Powered-By: PHP/4.1.2
    Transfer-Encoding: chunked
    Content-Type: text/html; charset=ISO-8859-1]]>
  </http-header>
</metadata>
```

The second HTTP request (getfile) should return the archived file with the given archive identifier (aid).

Compiling an aid-list for the Exporter

The list to feed the is an XML-file with the archive identifiers for the documents that you want exported. An example aid list is given below.

```
<aidlist>
  <aid>/disk1/webarchive/2003-12-01/1/113735474c69a4fc04e79f3b76f143eb</aid>
  <aid>/disk1/webarchive/2003-12-01/1/1caa603b3cd9ed0a91771bf0dec528a0</aid>
  <aid>/disk1/webarchive/2003-12-01/1/33429a6c42546ef814a8789691952723</aid>
  <aid>/disk1/webarchive/2003-12-01/1/35859a0609a14d2d03169dab44da6a2a</aid>
  <aid>/disk1/webarchive/2003-12-01/1/6ae46d2023c264ea5d88e1829bab9405</aid>
  <aid>/disk1/webarchive/2003-12-01/1/6dc9eb60d71b3b857c14d668c720a6a7</aid>
  <aid>/disk1/webarchive/2003-12-01/1/7f1a76c739a93a47e32993a37d5f879b</aid>
  <aid>/disk1/webarchive/2003-12-01/1/b5cc09e36e3ad276439747423d668f68</aid>
  <aid>/disk1/webarchive/2003-12-01/1/bd3ade5ed89630006dd92e6cc8d894d0</aid>
  <aid>/disk1/webarchive/2003-12-01/1/c106790e48c22e6533293131e3fa9832</aid>
  <aid>/disk1/webarchive/2003-12-01/1/d66fd294269c7ed04b2343fd87371dc9</aid>
  <aid>/disk1/webarchive/2003-12-01/1/d7fdd3a4ce3780ff5156c0166a79b4c0</aid>
</aidlist>
```

In order to do a first test of the Exporter you might want to make an aid-list manually containing just a few id's

To generate aid lists automatically you may use the script *gen_aids.pl* which you will find in the same directory as the Exporter. Run the script with the *-h* option for information on how to use the script.

Exporting

To start the the Export simply use

```
# perl exporter.pl -i <path><idfile> -o <outpath><filenameprefix>
```

where *path* is the path to the directory where the list of ids is stored, *idfile* is the name of the id-file, *outpath* is the path to the directory where you want the output to be stored and *filenameprefix* is a prefix for the files that the Exporter will output.

Indexing

In order to index the output from the Exporter using the NWA adapted Lucene Search Engine follow the instructions in <http://<tomcatserver>:<port>/nwa/index.html>, where *tomcatserver* and *port* is the hostname and portnumber of the tomcat server you have installed.

Searching

Open a browser and type in the URL <http://<installurl>/search2.php>, where *installurl* is the "URL for this installation" you typed in when installing the Access Module (see the chapter on installation).

Chapter 3. Configuring

The parameters set by the install dialogue should be sufficient for getting you up running. Below is given a list of files that has to be edited manually when changing the configuration. More in-detail information about the configuration of the NWA Toolset will be provided in a later version of the documentation

Exporter configuration

Configuring the Exporter is done by editing the *exporter.conf* file reciding in the conf directory of your Exporter installation

Access Module configuration

Configuring the Access Module is done by editing the files:

- *config.inc* reciding in the Access Module installation directory.
- *luceneconfig.inc* reciding in the directory *<installdir>/include/seal/lucene*.