

---

# NwaToolset Manual

Sverre Bang

Krister Persson

John Erik Halse

\$Id: manual.xml,v 1.1 2004/10/21 11:05:43 sverreb Exp \$

Copyright © 2003, 2004 Royal Library in StockholmRoyal Library in CopenhagenHelsinki  
University Library in FinlandNational Library of NorwayNational and University Library of  
Iceland

## Table of Contents

Introduction .....	1
Overview .....	2
Installation .....	6
Obtaining the NwaToolset .....	6
Installing .....	6
Using the NwaToolset .....	8
Testing the Retriever .....	8
Exporting .....	10
Indexing .....	11
Searching .....	11
Manual Configuration .....	11
Exporter configuration .....	11
Access Module configuration .....	11
Retriever Configuration .....	11
Step-by-step installation example .....	12
Installation dialogue .....	12
Configuration .....	14
Generating lists of AID's and testing the Retriever .....	14
Exporting .....	18
Indexing .....	19
Providing Access .....	21

## Introduction

The NWA Toolset is a freely available solution for searching and navigating archived web document collections.

The NWA (Nordic Web Archive) is the Nordic National Libraries' forum for co-ordination and exchange of experience in the fields of harvesting and archiving web documents. Since November 2000 the NWA cooperation has been developing the NwaToolset. The activity has been funded by Nordunet2, NORDINFO and the Nordic National Libraries. The NWA toolset was built using PHP, Perl and Java. It utilizes open standards like the http protocol and XML extensively for communication between different parts of the system. The actual use of the NWA toolset (i.e. searching and navigating a web archive) is done via a regular web browser, and no special browser plugins are needed to make it work.

A web archive may consist of a large number of web documents, but also several versions of the same web document (i.e. the documents where downloaded from the same URL). Potential users of the NWA Toolset might be anyone that has a web archive. Examples of such users may be:

- National Libraries or other organisations collecting parts of the internet for long term preservation.
- Companies or organisations keeping a historical collection of their own web site and/or intranet.
- Private persons keeping a historical collection of their own web site.

Note that in the following text a archived file and a web page is not necessarily the same thing. What the user experience as one web document may consist of several archived files (e.g. a web page which comprises the html file and the inline images).

## Overview

The NWA Toolset consists of four main parts. These are the Document Retriever, the Exporter, Access Module and the Search Engine.

- The Document Retriever serves as the interface to the web archive. It delivers archived files and associated metadata to the Exporter and the Access Module upon request.
- The purpose of the Exporter is to transform the archived files and its associated metadata to an intermediate XML format named the NWA Document Format. The NWA Document Formatted document collection is then fed to the indexer of a search engine.
- The Access Module interfaces both the search engine and the Document Retriever thus giving the user the possibility to search, browse and navigate the archived web documents.
- A search engine, including a search engine abstraction layer. At present, the NWA Toolset supports the Apache Jakarta Lucene search engine and the search engine from FAST Search & Transfer ASA. If the NWA Toolset, at the time of release does not support your search engine of choice, you will have to implement your own search engine abstraction layer.

In addition, a web archive is needed. A key requirement for the archive is that the files are stored unaltered and that a metadata set consisting of at least the original url and timestamp of the archived files is available.

## Exporter

The Exporter fetches archived files and associated metadata from the web archive and prepares them for indexing.

The input to the Exporter is a list of archive id's defining which archived files the Exporter should process. The list is generated at the archive side as a preparation for export. Automated tasks for creating such lists will have to be tailored to suit the specific web archive site.

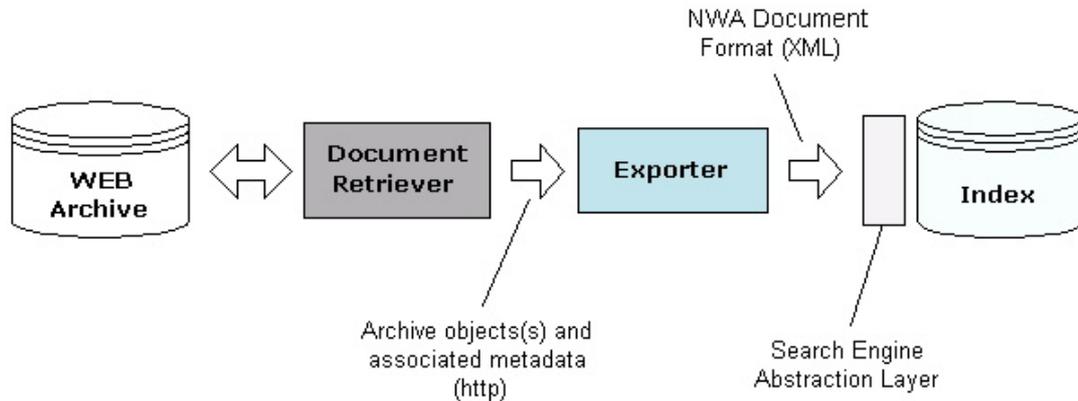
For each id in the list the following happens:

1. The Exporter requests, and receives metadata for the archived file.
2. If the archived file is of type html, the Exporter requests the file, receives the file and extracts data from the file.
3. The Exporter outputs the extracted data along with relevant metadata to the NWA document format.

If the file is not html (e.g. a gif-image) the only data exported for the file will be available metadata like its archive id, the original URL of the file, its mime type and its timestamp (e.g. time of harvesting).

The Exporter is prepared for interfacing a converter that transforms non-html text files like pdf, ms-word etc. into html thus enabling extraction of data from these files as well. It is also prepared for interfacing language detection software enabling the user to narrow a search to text-content files written in a specific language. These tools are not part of the NWA Toolset and they must therefore be obtained otherwise. In the NWA project, third party products licensed by FAST Search & Transfer ASA were used for both the html conversion and the language detection.

**Figure 1. Exporting**



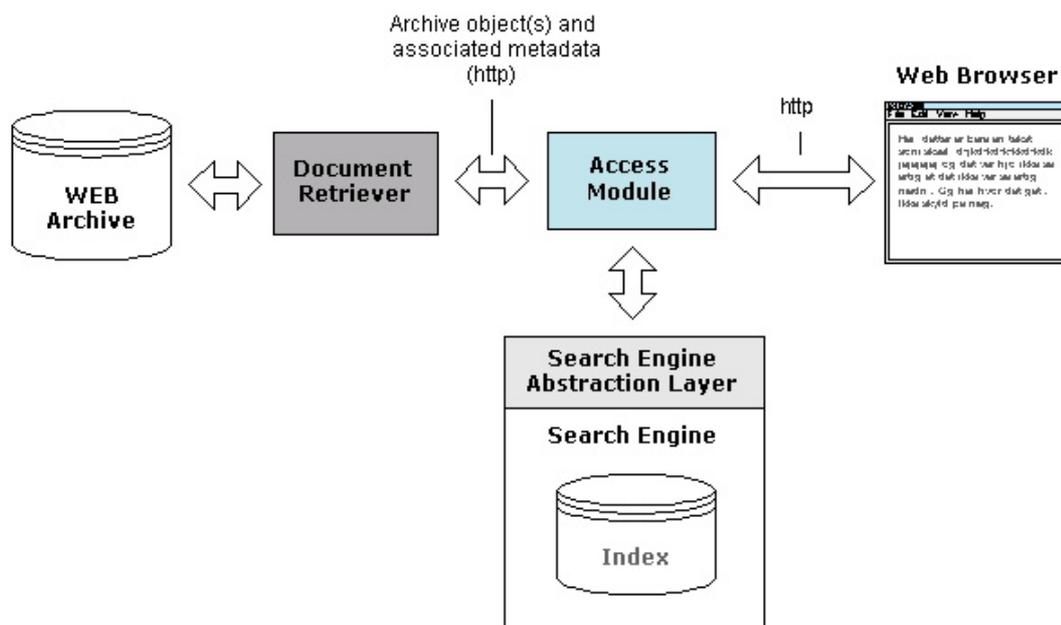
## Access Module

The Access Module provides the user with interfaces for searching, browsing and navigating the archived web pages.

When the user submits a query, the Access Module uses the search engine to find the archived files containing the text(s) satisfying the query. When the user asks for a specific URL the Access Module will return the archived file with that particular URL (e.g. the archived file originally downloaded from the url <http://www.nb.no/index.html>). Before the file is delivered to the user's browser the file is parsed and all the inline links and references are altered to point into the archive rather than out to the Internet. When the browser encounters inline references (e.g. an image), the browser will ask the Access Module to return these files in order to present them as part of the web page.

The resulting web page is presented with a timeline at the top and the web document below it. The timeline queries the index for all archived versions of the web page and displays the timestamps graphically along the line.

**Figure 2. Access**



Searching a web archive through the Access Module resembles using a Internet search engine like Google. An example of the NwaToolset search interface is shown below.

**Figure 3. Search result**

The screenshot shows the NWA search interface in Mozilla Firefox. The browser window title is 'NWA search - Mozilla Firefox'. The address bar shows the URL: `http://nwa.nb.no/demo/search.php?querytype=`. The search results are displayed in a table with three entries:

Match:	Query:	Search
All words	nasjonalbiblioteket	Search
Year (from - to)		
[ ] - [ ]		

Total number of versions found : **31**. Displaying web pages **1-3**

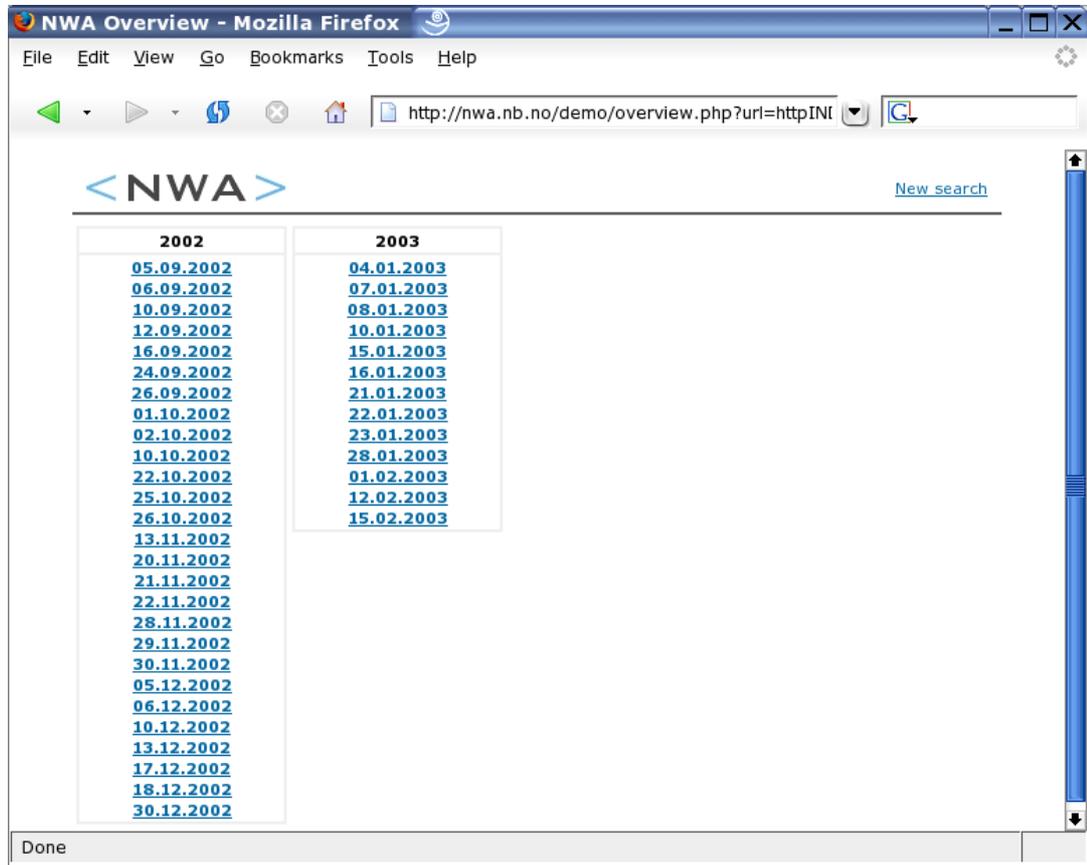
- ...av amanuensis **Svein Sundbø** i **Nasjonalbiblioteket** innholder et...nb.no/nati Ny **Dewey**-utgave (DDK5) utgitt av **Nasjonalbiblioteket** Et stort arbeid...av ny norsk **Dewey**-utgave - pressemelding **Nasjonalbiblioteket** årsmelding...(http://www.nb.no/index.html)  
Number of versions satisfying query / total number of versions : 21/40  
[Timeline](#) | [Overview](#)
- ...bestilles fra **Nasjonalbiblioteket**. 31.10.02 Ny diskusjonsliste...administreres av **Nasjonalbiblioteket**. Tanken er at...av amanuensis **Svein Sundbø** i **Nasjonalbiblioteket** innholder et fylldig...Lansering av ny norsk **Dewey**-utgave 08.08.02...(http://www.nb.no/html/nyheter.html)  
Number of versions satisfying query / total number of versions : 8/16  
[Timeline](#) | [Overview](#)
- ...29.11.02 Ibsenbrev til **Nasjonalbiblioteket** **Nasjonalbiblioteket** sikret seg fire brev...utarbeidet av amanuensis **Svein Sundbø** i **Nasjonalbiblioteket** innholder et fylldig forord...Lansering av ny norsk **Dewey**-utgave 08.08.02 Billedsamlingen...(http://www.nb.no/html/nyhetsarkiv\_2002.html)  
Number of versions satisfying query / total number of versions : 2/2  
[Timeline](#) | [Overview](#)

Results: 1-3 |

Done

Clicking the Overview link of a specific hit will display all the dates for the versions found for the chosen URL (the overview does not contain any information of which versions that actually satisfied the query term given in the first place).

**Figure 4. Overview**



Clicking one of the links in the Overview page will take the user to the Timeline view with the chosen version displayed. Clicking the Timeline link in the Search result page will also take the user to the Timeline page but the version displayed will be the most recent of the versions satisfying the query term given.

**Figure 5. Timeline**



The Timeline view shows the different versions of a given URL displayed graphically along a timeline. The actual archived web page is shown below the Timeline. All links and inline references are altered before the web page is transmitted to the user (i.e. the users browser).

When navigating from the Overview or the result list of a search interface, the URL of the chosen version is passed along and shown in the URL field. A URL may also be entered manually.

Navigation between the different versions is done by directly clicking a specific point on the timeline, or by using the arrows first, previous, next and last.

When entering the timeline view the resolution is set to auto. This means that the timeline automatically drills down to the resolution needed to display single versions along the line. The Auto checkbox may be unchecked in order to manually choose the resolution (choosing a different resolution when in auto also disables auto resolution).

It is also possible to perform a simple search from the Timeline by typing in a query term and pressing *Go*.

## Installation

This chapter describes how to obtain, install and configure the NwaToolset..

## Obtaining the NwaToolset

The latest version of the NwaToolset may be downloaded from the NwaToolset home page [http://nwatoolset.sourceforge.net] at sourceforge. If you don't have a Web Archive readily available you may download the sample archive as well.

## Installing

To install the NwaToolset do the following:

- Unpack `nwatoolset_<buildno>.tar.gz` in a temporary directory
- Invoke the installer using `./install_nwatoolset.pl`
- Follow the on-screen instructions

If you want to distribute the different components of the NwaToolset on different hosts, you need to repeat parts of the installation on the other hosts.

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

The installation steps are shown below.

- Install a 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 indexer), the presence of a working Retriever is an absolute requirement. Currently the NWA Toolset supports Nedlib based and Heritrix-based (ARC) archives.
- 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 (XML).
- 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 NWA Toolset include the Apache Jakarta Lucene search engine adapted for NWA Toolset use.

## System Requirements

The NWA Toolset and the NWA Adapted Lucene search engine has been tested on different builds of *RedHat* (7.3, 8, AS2 etc.), *Fedora* and *Suse* Linux. There is no reason to believe that the system will not work on other linux/unix ditributions.

## Installer Requirements

In order to run the install script a working version of *Perl* has to be installed (Even if the installer may be invoked using any version of Perl, the Exporter requires a specific Perl version).

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

## Retriever Requirements

The Retriever for Nedlib based 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/>

The Retriever for Heritrix generated web archives requires a java servlet container like Apache Jakarta Tomcat.

## 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).
- A Java servlet container like Apache Jakarta Tomcat.

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>

# Using the NwaToolset

After installing the NwaToolset you should go through the following steps.

1. Test the that the Retriever is functioning correctly
2. Compile a list of files to export and execute an export based on that list
3. Index the output from the Exporter and verify that a vaild index is produced.
4. Check the search and navigate functionality

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]/<retriever>?reqtype=<reqtype>&aid=<aid>`

Where *retriever* is the retriever script doing the retrieval, *reqtype* is the request type and the *aid* is the unique identifier (within the archive) given to the harvested file. The *getmeta* request will return archived technical metadata for the file in question and the *getfile* request will return the archived file itself.

An example of a reply to a getmeta request is given below.

```
<?xml version="1.0" encoding="UTF-8" ?>
<retrievermessage>
  <head>
    <reqtype>getmeta</reqtype>
    <aid>
      /var/nwadata/nedlib/nbvev/2002-10-10/1/5d4cfd51bd7952a163fbae25fbe8eb2c
    </aid>
  </head>
  <body>
    <metadata>
      <status>online</status>
      <content_length>2581</content_length>
      <content_checksum>3be2e91f1765183742836919cca282e3
      </content_checksum>
      <last_modified_time>20021009081522</last_modified_time>
      <contenttype>
        <type>text/html</type>
        <charset></charset>
      </contenttype>
      <archival_time>20021010003212</archival_time>
      <url><![CDATA[http://www.nb.no:80/baser/kunstbib]]></url>
      <http-header><![CDATA[HTTP/1.1 200 OK
        Date: Thu, 10 Oct 2002 00:32:05 GMT
        Server: Apache/1.3.26 (Unix) PHP/4.1.2
        Last-Modified: Wed, 09 Oct 2002 08:15:22 GMT
        ETag: "19896-a15-3da3e59a"
        Accept-Ranges: bytes
        Content-Length: 2581
        Content-Type: text/html ]]
      </http-header>
    </metadata>
  </body>
</retrievermessage>
```

## Testing the Nedlib Retriever

Before testing the Retriever you will need to edit its configuration file. See the section on manual configuration for details.

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

- `ht-tp://<hostname>.<domainname>[:port]/<retrieverpath>/docretriver_nedlib.php?reqtype=getmeta&aid=<archivepath>/2002-12-01/1/ebec0183244bb27d77989efc5be91218`
- `ht-tp://<hostname>.<domainname>[:port]/<retrieverpath>/docretriver_nedlib.php?reqtype=getfile&aid=<archivepath>/2002-12-01/1/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.

## Testing the ARC Retriever

In order to test the ARC Retriever open the following URL in a browser:

`http://<hostname>:<port>/ArcRetriever/`

where *hostname* is the name of your host and *port* is the port where your Java Servlet Container is running and do the following:

1. Generate an aid file by following the instructions in the web interface above (you will need to have at least one ARC-file available on disk before doing this).
2. Copy one of the AID's from the generated AID-list.
3. In your browser open the URLS :
  - `http://<hostname>:<port>/ArcRetriever/ArcRetriever.jsp?reqtype=getmeta&aid=<aid>`
  - `http://<hostname>:<port>/ArcRetriever/ArcRetriever.jsp?reqtype=getfile&aid=<aid>`

## Exporting

To start the Export simply invoke the Exporter from the command line using:

```
# 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 file containing the AID's to export, *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.

The list to feed the Exporter is an XML-file with the AID's of the archived files you want exported. An example of an AID list generated by the ARC Retriever is given below (see the section above on testing the ARC Retriever for details on this).

```
<aidlist>
<aid>675//var/nwadata/arcs/IAH-20040928123615-00047-utvikling1.nb.no.arc.gz</aid>
<aid>3607//var/nwadata/arcs/IAH-20040928123615-00047-utvikling1.nb.no.arc.gz</aid>
<aid>4681//var/nwadata/arcs/IAH-20040928123615-00047-utvikling1.nb.no.arc.gz</aid>
<aid>5214//var/nwadata/arcs/IAH-20040928123615-00047-utvikling1.nb.no.arc.gz</aid>
<aid>5612//var/nwadata/arcs/IAH-20040928123615-00047-utvikling1.nb.no.arc.gz</aid>
..
.
</aidlist>
```

Below is shown an AID list for a Nedlib generated archive.

```
<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>
..
.
</aidlist>
```

In order to generate an AID list for a Nedlib based archive you need to run the Perl script `generate_identifiers.pl`. Please note that the script is not part of the Nedlib Retriever install but the Exporter install (the script is installed alongside the `exporter.pl` script). Usage :

```
# perl generate_identifiers.pl -o <output-prefix> [-r] [-d  
<inputdirectory>] [-n <number_of_ids_per_file>]
```

where:

-d <directory>: Specify input directory where to find archived files. Default is current working directory,

-r : Generate AID's recursively i.e. to traverse all sub directories of the directory given by -d,

-h : Help,

-o <output-prefix>: Specify output directory and filename-prefix. First outputfile has suffix -0, then -1 and so on.

-n <number\_of\_ids\_per\_file>: Specify the maximum number of identifiers in each output file. Default is 100.000

## Indexing

In order to index the output from the Exporter using the NWA adapted Lucene Search Engine follow the instructions in <http://<host>:<port>/nwa/index.html>, where *host* and *port* is the host name and port number of the Java Servlet Container you have installed (e.g Apache Jakarta Tomcat).

## Searching

Open a browser and type in the URL <installurl>/search.php, where *installurl* is the “URL for this installation” you typed in when installing the Access Module.

## Manual Configuration

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. If you want the Exporter to fetch from a different Retriever than given in the installation dialogue update the value of *retriever\_url* in *exporter.conf*.

## 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.

Please note that HTML parser of the Access Module is written in Perl and you may need to re-configure your Apache server somewhat to allow Perl scripts to be executed from the directory where it was installed.

## Retriever Configuration

Configuring the Nedlib Retriever is done by editing the *file conf\_retriever\_nedlib.inc*. The file re-

cides in the same directory as the retriever itself. The parameter `$conf_archive_directory` in the configuration file allows you to control what part of the file system the Retriever has access to.

## Step-by-step installation example

The information in this section was recorded from an actual installation done. The steps described here ended up in a fully functional NwaToolset. All the steps were done both for Nedlib and ARC based archives, but for those steps involving the same operations for Nedlib and ARC only one example is given.

## Installation dialogue

The below screendump shows the installation dialogue. Some extra white space has been inserted and user input is shown in bold face to increase readability.

```

sverreb@sverreb:~/test> ll
total 2547
-rw-r--r--  1 sverreb users 2602877 2004-10-18 21:32 nwatoolset_1_1_RC5.tar.gz

sverreb@sverreb:~/test> tar xvfz nwatoolset_1_1_RC5.tar.gz
nwatoolset_1_1_RC5/
nwatoolset_1_1_RC5/doc/
nwatoolset_1_1_RC5/retriever.tar
nwatoolset_1_1_RC5/install_nwatoolset.pl
nwatoolset_1_1_RC5/ArcRetriever.war
nwatoolset_1_1_RC5/accessmodule.tar
nwatoolset_1_1_RC5/nwa.war
nwatoolset_1_1_RC5/exporter.tar
nwatoolset_1_1_RC5/utils.tar
sverreb@sverreb:~/test> cd nwatoolset_1_1_RC5/
sverreb@sverreb:~/test/nwatoolset_1_1_RC5> <emphasis
role="bold"
>ll</emphasis
>
total 3010
-rw-r--r--  1 sverreb users  358400 2004-10-18 21:32 accessmodule.tar
-rw-r--r--  1 sverreb users 2069350 2004-10-18 21:32 ArcRetriever.war
drwxr-xr-x  2 sverreb users     48 2004-10-18 21:32 doc
-rw-r--r--  1 sverreb users 194560 2004-10-18 21:32 exporter.tar
-rwxr-xr-x  1 sverreb users  19009 2004-10-18 21:32 install_nwatoolset.pl
-rw-r--r--  1 sverreb users 395625 2004-10-18 21:32 nwa.war
-rw-r--r--  1 sverreb users  20480 2004-10-18 21:32 retriever.tar
-rw-r--r--  1 sverreb users  10240 2004-10-18 21:32 utils.tar

sverreb@sverreb:~/test/nwatoolset_1_1_RC5> ./install_nwatoolset.pl

Type in your host name or leave blank for default value, default is
[sverreb]:sverreb.nb.no

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

```
<emphasis
role="bold"
>1 2 3 4</emphasis
>
```

#### DOCUMENT RETRIEVER INSTALLATION

Type in numbers of components you want to install:

1. Nedlib Retriever
2. ARC Retriever

1 2

#### 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.

The Retriever should be installed into a directory where you can execute scripts.

Type in the name of a suitable directory.

The Retriever will be installed in the directory 'nwatoolset/retriever/' relative to the given directory.

Default is [/var/www/html]:/srv/www/htdocs

Creating directory /srv/www/htdocs/nwatoolset/retriever

Nedlib Retriever has been installed into /srv/www/htdocs/nwatoolset/retriever. Please note that the retriever has to be configured manually. See documentation for details.

Finished installation of Retriever

#### STARTING INSTALLATION OF ARC RETRIEVER

This module needs the existence of a servlet container (eg. Jakarta Tomcat). Type in the full path to where 'war' files should be deployed for your container. For Tomcat this will typically be \$Tomcat\_install\_dir/webapps.

Type in full path to where 'war' files should be deployed or leave blank for default value, default is []:/usr/share/tomcat/webapps

Copied the Nwa Arc Retriever Web application to /usr/share/tomcat/webapps

#### CHECKING MODULES IN YOUR PERL INSTALLATION

Writing Makefile for Nwatoolset

#### 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/nwatoolset]:

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://sverreb.nb.no/nwatoolset/retriever/retriever_nedlib.php]:`

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

Type in the name of your collection default is `[test]:no test1`  
Finished installation of Exporter files. They were placed into `/usr/local/nwatoolset/bin` and `/usr/local/nwatoolset/conf`

The Exporter's log file will be placed in the directory `/usr/local/nwatoolset/log`

#### 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]:/srv/www/htdocs`

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

Type in the name of your collection (same as used in Exporter installation), default is `[test]:no test`

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

Installed Access module into `/srv/www/htdocs/nwatoolset`

#### STARTING INSTALLATION OF LUCENE SEARCH ENGINE

Copied the Nwa Lucene Web application to `/usr/share/tomcat/webapps`

`sverreb@sverreb:~/test/nwatoolset_1_1_RC5>`

## Configuration

### Nedlib Retriever

The configuration file, `conf_retriever_nedlib.inc` was updated so that the Retriever only could access the directory where the Nedlib archive actually resided.

```
$conf_archive_directory = "/var/nwadata/nedlib/";
```

## Generating lists of AID's and testing the Retriever

### Nedlib based Archive

The screen listing below shows (abbreviated) the output from the AID generator for Nedlib based archives.

```
sverreb@sverreb:~> cd /usr/local/nwatoolset/bin
sverreb@sverreb:/usr/local/nwatoolset/bin> ./generate_identifiers.pl -o
/var/nwadata/nedlibaids/idlist -r -d /var/nwadata/nedlib/nbvev/

The maximum number of identifiers in each outputfile set to 100.000

Working with: /var/nwadata/nedlib/nbvev
Working with: /var/nwadata/nedlib/nbvev/2002-10-10
Working with: /var/nwadata/nedlib/nbvev/2002-10-10/1
Working with: /var/nwadata/nedlib/nbvev/2002-10-11
Working with: /var/nwadata/nedlib/nbvev/2002-10-11/1
..
.
Working with: /var/nwadata/nedlib/nbvev/2003-2-7/1
Working with: /var/nwadata/nedlib/nbvev/2003-2-8
Working with: /var/nwadata/nedlib/nbvev/2003-2-8/1
Working with: /var/nwadata/nedlib/nbvev/2003-2-9
Working with: /var/nwadata/nedlib/nbvev/2003-2-9/1

Finished writing 8586 IDs to AID-file: /var/nwadata/nedlibaids/idlist-0

Found 8586 documents in total.

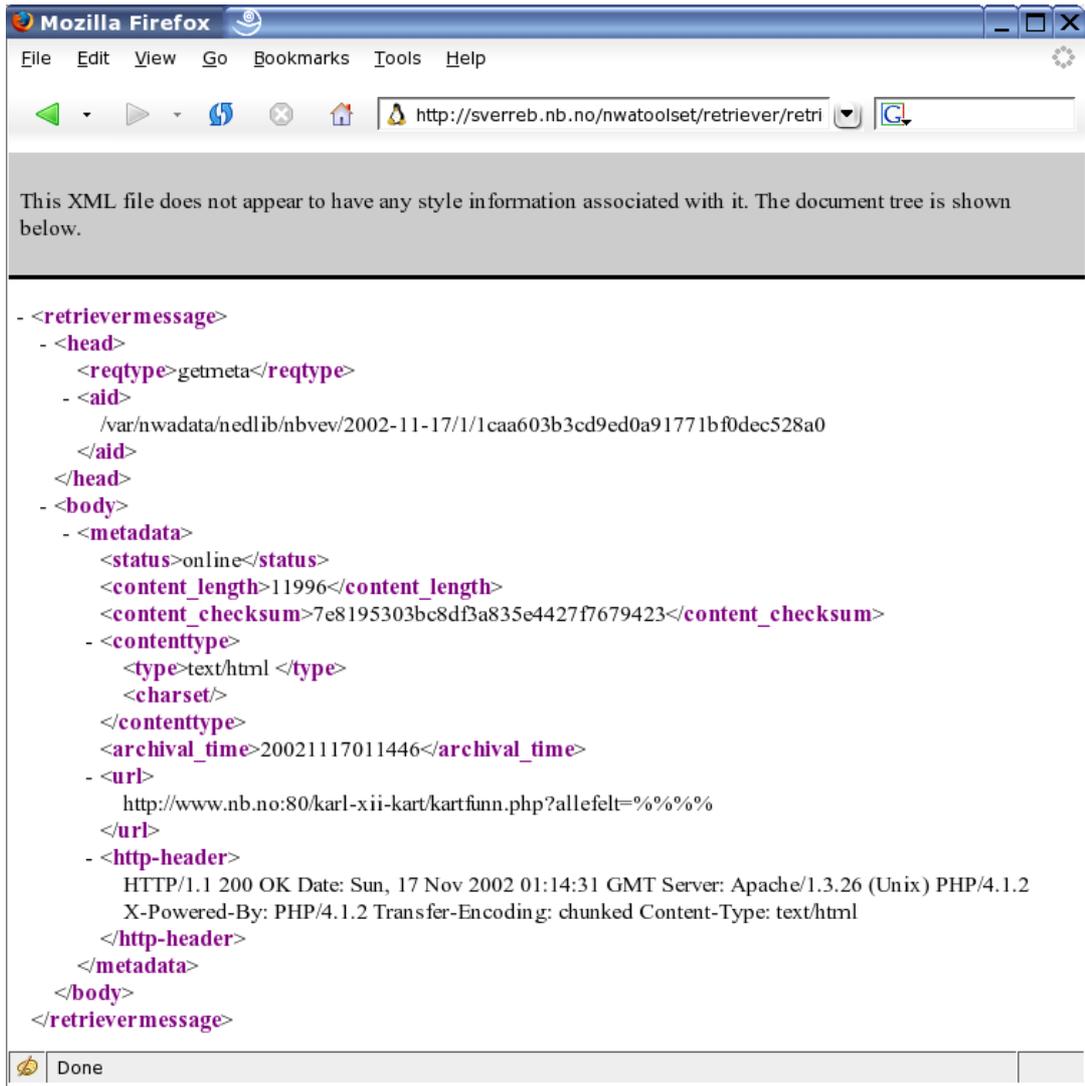
Done!

sverreb@sverreb:/usr/local/nwatoolset/bin>
```

A random AID was picked from the list of AID's generated above. The AID was entered into the browser's URL field as follows:

```
http://sverreb.nb.no/nwatoolset/retriever/retriever_nedlib.php?aid=
/var/nwadata/nedlib/nbvev/2002-11-17/1/1caa603b3cd9ed0a91771bf0dec528a0&reqtype
```

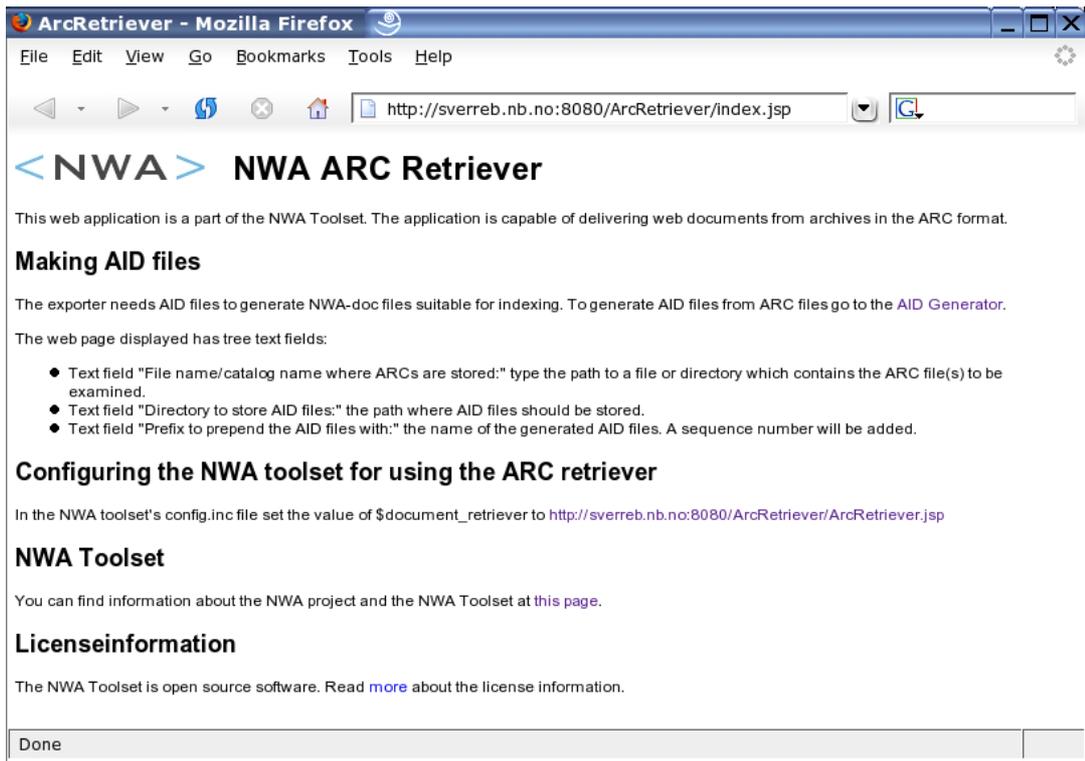
**Figure 6. Metadata from Nedlib Retriever**



## ARC based archive

Tomcat were restarted so that the installed ArcRetriever.war file were deployed by Tomcat. The ARC retriever now was available at the URL: <http://sverreb.nb.no:8080/ArcRetriever/>. See below for screenshot.

**Figure 7. Arc Retriever start page**



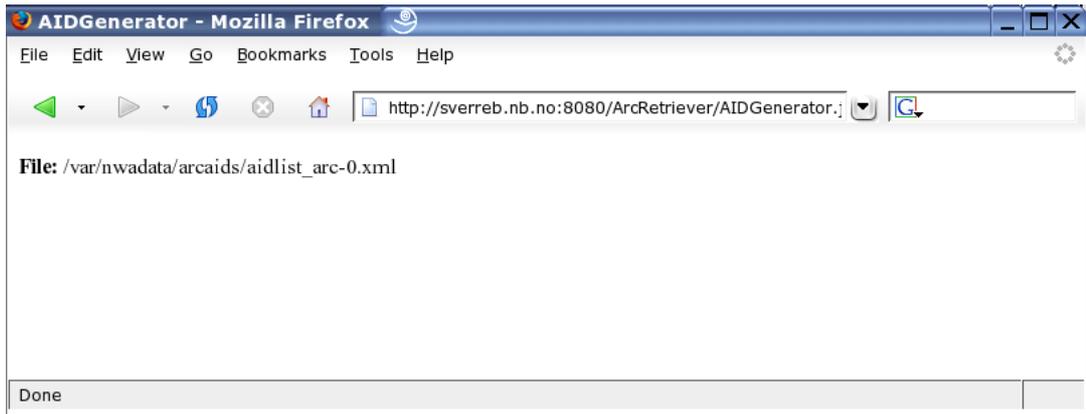
The link to the AID generator were clicked. The missing data needed were filled in (see screenshot below).

**Figure 8. AID Generator**



After submitting the info in the previous screenshot the following result emerged:

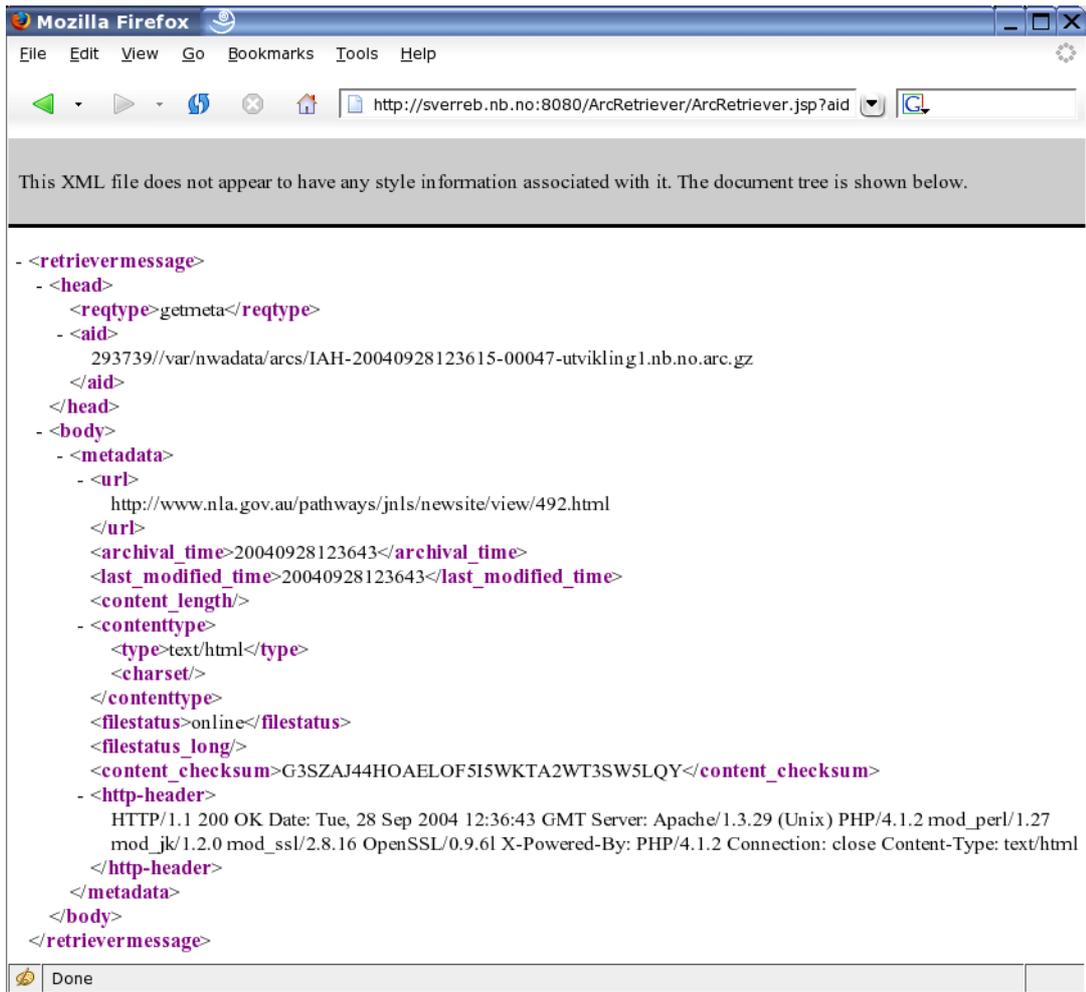
**Figure 9. AID Generator done**



A random AID was picked from the list of AID's generated above. The AID was entered into the browser's URL field as follows:

```
http://sverreb.nb.no:8080/ArcRetriever/ArcRetriever.jsp?aid=
293739//var/nwadata/arcs/IAH-20040928123615-00047-utvikling1.nb.no.arc.gz&reqty
```

**Figure 10. Metadata from ARC Retriever**



## Exporting

An export from a Nedlib based archive were done using the following command

```
./exporter.pl -i /var/nwadata/nedlibaids/idlist-0 -o /var/nwadata/nwadoocs/from_nedlib/nwadoocs
```

To export the ARC based archive the exporter.conf file had to be updated. The parameter set retriever\_url was changed to http://sverreb.nb.no:8080/ArcRetriever/ArcRetriever.jsp (because the retriever\_url set during installation pointed to the Nedlib Retriever. See installation dialogue earlier in this document.).

The ARC Export were started using:

```
./exporter.pl -i /var/nwadata/arcaids/aidlist_arc-0.xml -o /var/nwadata/nwadoocs/from_arc/nwadoc
```

The screen output (abbreviated) from the ARC Export is shown below.

```
sverreb@sverreb:/usr/local/nwatoolset/bin>./exporter.pl -i /var/nwadata/arcaids/aidlist_arc-0.xml -o /var/nwadata/nwadoocs/from_arc/nwadoc
```

```
Warning: Perl is no longer using LC_TYPE 'en_GB.UTF-8', but LC_TYPE 'C'
Reading identifiers from input file: /var/nwadata/arcaids/aidlist_arc-0.xml
```

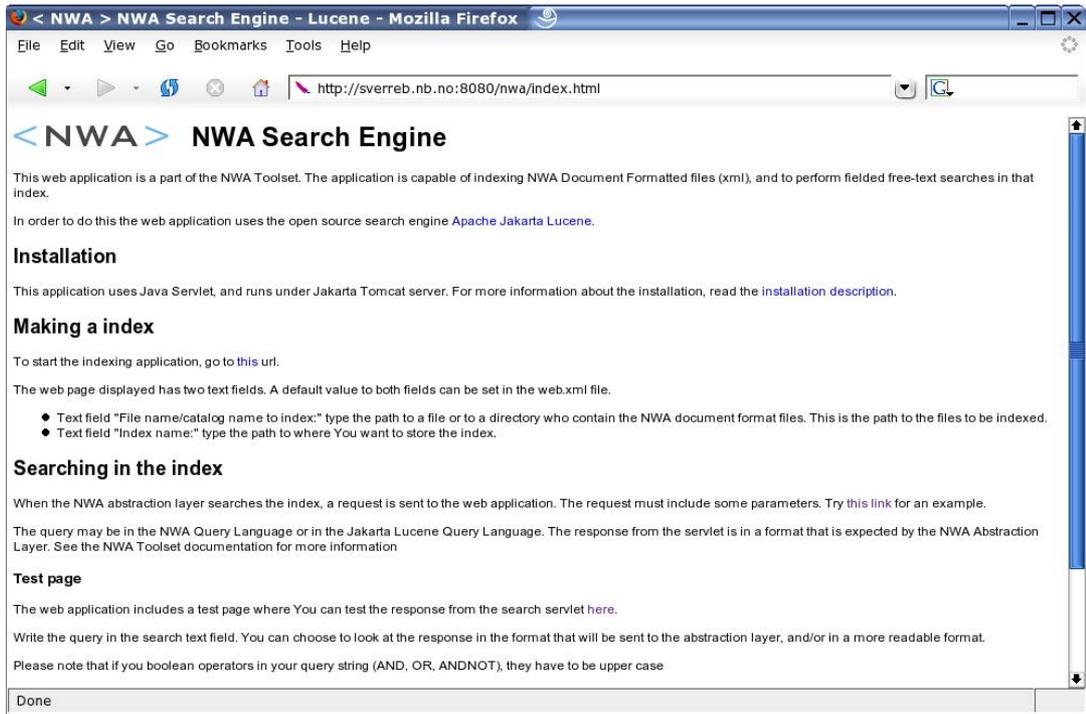
```
2925 identifiers found.
```

```
exporter.pl[INFO] starting at Tue Oct 19 11:16:28 2004.
child 1472: 2925 remaining documents (100.00% remaining)
Running Childs[child 1472]: 1( of max 10 children)
child 1472: 2425 remaining documents (82.91% remaining)
Running Childs[child 1472]: 2( of max 10 children)
child 1472: 1925 remaining documents (65.81% remaining)
Running Childs[child 1472]: 3( of max 10 children)
child 1472: 1425 remaining documents (48.72% remaining)
Running Childs[child 1472]: 4( of max 10 children)
child 1472: 925 remaining documents (31.62% remaining)
Running Childs[child 1472]: 5( of max 10 children)
child 1472: 425 remaining documents (14.53% remaining)
Running Childs[child 1472]: 6( of max 10 children)
Child 1488: 50/500 identifiers (10 %) finished
Child 1518: 50/500 identifiers (10 %) finished
Child 1493: 450/500 identifiers (90 %) finished
Child 1518: 500/500 identifiers (100 %) finished
Child 1518 [runChild] Finished the processing of 500 documents
..
..
Child 1498: 450/500 identifiers (90 %) finished
Child 1504: 500/500 identifiers (100 %) finished
Child 1524: 420/425 identifiers (100 %) finished
Child 1504 [runChild] Finished the processing of 500 documents
Child 1524 [runChild] Finished the processing of 425 documents
Child 1493: 500/500 identifiers (100 %) finished
Child 1493 [runChild] Finished the processing of 500 documents
Child 1498: 500/500 identifiers (100 %) finished
Child 1498 [runChild] Finished the processing of 500 documents
exporter.pl[INFO] finished at Tue Oct 19 11:20:27 2004.
sverreb@sverreb:/usr/local/nwatoolset/bin>
```

## Indexing

The Indexer Web Application were accessed through a browser using: **http://sverreb.nb.no:8080/nwa/**

**Figure 11. Indexer Web Application Start page**



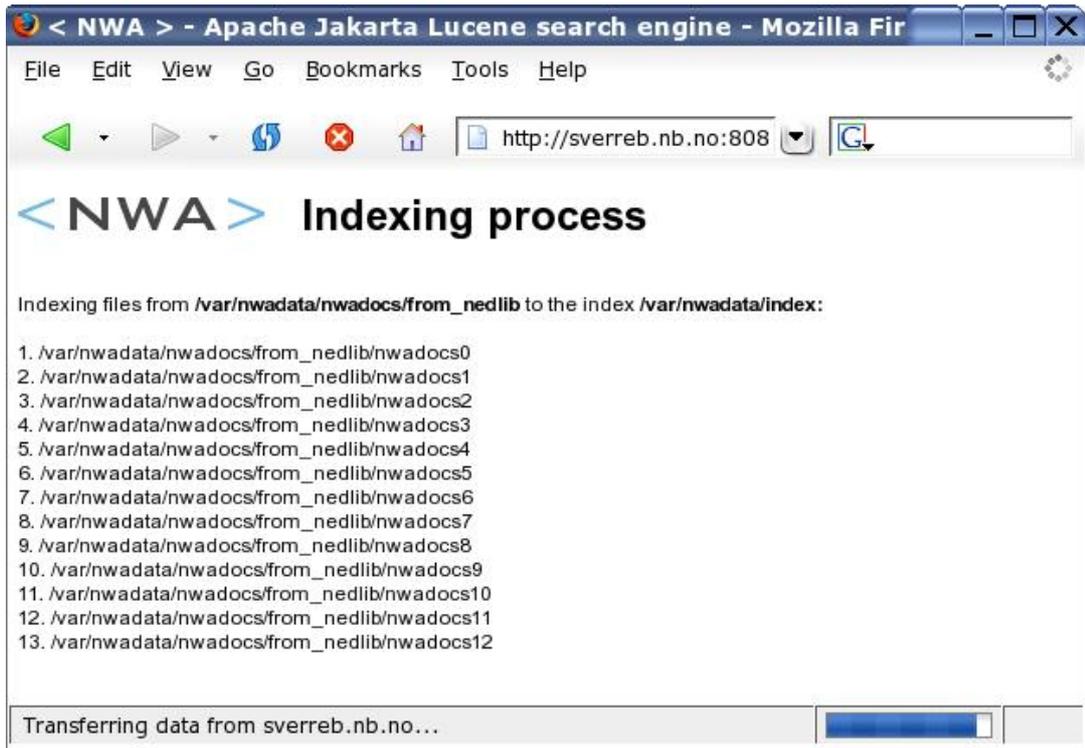
Indexing preparations were made:

**Figure 12. Starting the Indexer**



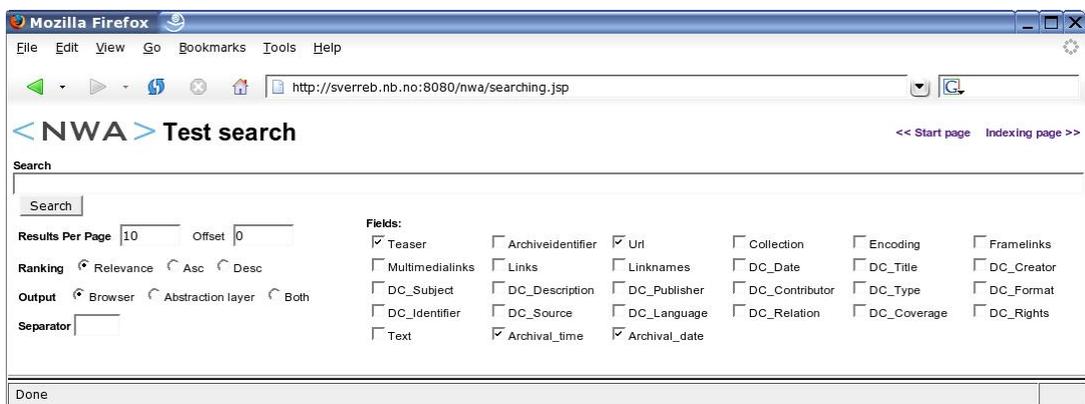
Indexing were started:.

**Figure 13. Indexing Process**



The indexer was left until the Web Application reported done. From the indexer start page the test search interface was launched.

**Figure 14. Test Search Interface**



Through the search interface it is possible to control what the result should contain and how it should be sorted. The query submitted has to comply with NWA Query Language. Boolean operators has to be given in upper case.

## Providing Access

The Access Module was entered through the URL <http://sverreb.nb.no/nwatoolset/>.