NetarchiveSuite 5.2.x Release Notes

5.2.2 Release Date 25th November 2016

5.2.1 Release Date 23rd November 2016

5.2 Release Date: 4th November 2016

Contents

Most-recent updates for 5.2.x:

- **Download NetarchiveSuite**
- **Download Heritrix 3 Bundle (required)**
- **Javadoc**
- **Manuals**

Highlights in 5.2.2

NAS 5.2.2 restores the functionality, missing since the upgrade to Heritrix 3, which allows one to switch deduplication on or off as a setting to the `HarvestJobManager` component. The setting in `settings_HarvestJobManagerApplication.xml` is `harvester.harvesting.deduplication.enabled` which is binary valued. The setting is applied to harvests generated using any crawler template which includes the DeDuplicator bean, and which specifies the appropriate placeholder, for example as follows:

```
<bean id="DeDuplicator"
class="is.hi.bok.deduplicator.DeDuplicator">
    <property name="indexLocation"
value="%{DEDUPLICATION_INDEX_LOCATION_PLACEHOLD
ER}"/>
    <property name="matchingMethod"
value="URL"/>
    <property name="tryEquivalent"
value="TRUE"/>
    <property name="changeContentSize"
value="false"/>
    <property name="mimeFilter"
value="^text/.*"/>
    <property name="filterMode"
value="BLACKLIST"/>
    <property name="origin" value=""/>
    <property name="originHandling"
value="INDEX"/>
    <property name="statsPerHost"
value="true"/>
    <property name="enabled"
value="%{DEDUPLICATION_ENABLED_PLACEHOLDER}" />
  </bean>
```

The %{DEDUPLICATION_ENABLED_PLACEHOLDER} is replaced with the current value of the setting when jobs are generated. The placeholder is optional, and deduplication will be enabled by default for any template which includes the DeDuplicator in its disposition chain, and for which the "enabled" property is not explicitly defined.

Highlights in 5.2.1

NAS 5.2.1 is a bugfix release addressing in issue in wayback-indexing of deduplicate records.

Highlights in 5.2

Java 8

NetarchiveSuite now requires a Java 8 runtime for all components.

New Settings

- ChecksumFileApplication

```
/**
*
<b>settings.archive.checksum.usePrecompute
dChecksum</b>: This decides whether or not
use the pre-computed checksum sent as part
of the StoreMessage and UploadMessage
* The default is false
*/
    public static String
CHECKSUM_USE_PRECOMPUTED_CHECKSUM_DURING_U
PLOAD=
"settings.archive.checksum.usePrecomputedC
hecksumDuringUpload";
```

This boolean can be used to optimise the upload process to the bitarchives.

- GUIApplication, HarvestJobManager

```
/**
 *
<b>settings.common.topLevelDomains.tld</b>
: <br>
 * Extra valid top level domain, like
.co.uk, .dk, .org., not part of current
embedded public_suffix_list.dat file
 * in
common/common-core/src/main/resources/dk/n
etarkivet/common/utils/public_suffix_list.
dat
 * downloaded from
https://www.publicsuffix.org/list/public_s
uffix_list.dat
 */
public static String TLDS =
"settings.common.topLevelDomains.tld";
```

- HarvestControllerApplication

```
/**
 * The version number which goes in
metadata file names like
12345-metadata-&lt;version
number&gt;.warc.gz
 */
public static String
METADATA_FILE_VERSION_NUMBER =
"settings.harvester.harvesting.metadata.fi
lename.versionnumber";
```

This parameter allows for the definition of different generations of metadata file.

```
/**
 *
<b>settings.harvester.harvesting.metadata.
compression</b> Do we compress the
 * metadata associated with a given
harvest job.
 * default: false
 */
public static String METADATA_COMPRESSION
=
"settings.harvester.harvesting.metadata.co
mpression";
```

Controls whether metadata files are generated in compressed (warc.gz) format.

- ViewerproxyApplication, IndexServerApplication, WaybackIndexerApplication

```
/**
 * Specifies the suffix of a regex which
can identify valid metadata files by job
number. Thus preceding
 * the value of this setting with .* will
find all metadata files.
 */
public static String
METADATAFILE_REGEX_SUFFIX =
"settings.common.metadata.fileregexsuffix"
;
```

This parameter allows one to determine which metadata files to include in indexing (for Viewerproxy or Wayback). The full regex string to be searched consists of the string <jobid>-<harvestid> followed by this suffix. The default value is `-metadata-[0-9]+.(w)?arc(.gz)?` which matches all metadata files using the standard NetarchiveSuite naming scheme.

- GUIApplication

```
/**
     *
<b>settings.harvester.viewerproxy.allowFil
eDownloads</b> If set to false, there will
be no links to
     * allow download of warcfiles via the
Viewerproxy GUI.
     */
    public static String
ALLOW_FILE_DOWNLOADS =
"settings.harvester.viewerproxy.allowFileD
ownloads";
```

A simple security feature to hinder operators from easily downloading harvested archive files. (default: true)

```
public static String
HERITRIX3_MONITOR_TEMP_PATH =
"settings.harvester.harvesting.monitor.tem
pPath";
```

Path to a directory which the new Heritrix3 monitor feature can use for caching. This is empty by default, and falls back to the system-wide temporary directory (usually /tmp).

Control Heritrix from NetarchiveSuite (beta)

In earlier versions of NetarchiveSuite, there was limited monitoring of running heritrix harvests in the NetarchiveSuite GUI, but management of running jobs required opening the Heritrix3 console itself. From NetarchiveSuite 5.2, much of the Heritrix3 console functionality has been moved into NetarchiveSuite. It is now possible, from NAS itself to:

- pause, unpause or terminate running heritrix jobs
- to inspect reports on running jobs
- to show the crawl-log of a running job, either in entirety or filtered by regex
- to show and manipulate the Heritrix frontier

These extensive new features are experimental in NAS 5.2 and the developers welcome feedback, bug-reports, and code-patches.

Top-Level Domains Can Be Defined Externally

From NAS 5.2, all ICANN-recognized domains are recognized as valid in NAS. NAS contains an embedded copy of https://publicsuffix.org/list/public_suffix_list.dat, but this may be overridden, if necessary, by placing an alternative copy at the hard-coded path conf/public_suffix_list.dat in the installation on the machine where the GUIApplication and HarvestJobManager run.

warc.gz metadata files

NAS now supports compression of metadata files (warc.gz format) via the setting `settings.harvester.harvesting.metadata.compression`.

Warc Revisit Records

NAS now generates WARC revisit records when using the `is.hi.bok.deduplicator.DeDuplicator` deduplicator.

Tomcat

The web GUI now uses an embedded tomcat, rather than Jetty, as a servlet container. This changeover should be invisible to the end user.

New Heritrix Version

NAS now uses the most recent (unofficial) Heritrix release from Kristinn Sigurðsson at the National Library of Iceland (version 3.3.0-LBS-2016-02).

RSS Crawling

The heritrix crawl-rss extension from Kristinn Sigurðsson at the National Library of Iceland now also comes bundles with NAS, and is therefore available for use in NAS crawls. (See RSS Harvests for documentation).

GUI Styling

The styling of the web interface has been improved.

Issues resolved in release 5.2.2

T	Summary	Status
	Deduplication via settings doesn't work	RESOLVED
	Bring groovy script into NAS	CLOSED

2 issues

Issues resolved in release 5.2.1

T	Summary	Status
	DeduplicateToCDXAdapter fails to identify new dedup format	CLOSED

1 issue

Issues resolved in release 5.2

T	Summary	Status
	Extend SurtPrefixedDecideRule to rewrite certain seeds after they have been converted to SURTs	RESOLVED
	Running job details pages (Harveststatus-running-jobdetails.jsp) are empty	RESOLVED
	Upgrade Jetty 6 to Jetty9	CLOSED
	Generate revisit records	CLOSED
	H1 harvester-controller should fail gracely, if it receives a H3 template	CLOSED
	Enable H3 Scripting calls from NAS GUI	CLOSED
	Upgrade to latests H3 from Iceland	RESOLVED
	Remove tlds from settings	RESOLVED
	Bring groovy script into NAS	CLOSED
	Support WARC Revisit records	CLOSED
	Write gzipped-warcs	RESOLVED
	Link to Heritrix host takes too long to appear	RESOLVED
	Replace Jetty with Tomcat in GUI	CLOSED
	Add precomputed checksum to allow higher upload throughput to the checksum replica	RESOLVED
	Make warc-download hideable	RESOLVED
	Plugin of Domain definition suggested	RESOLVED
	Empty domain crawlertraps are inserted into H3 templates	RESOLVED

	History/Harveststatus-running-jobdetails.jsp?jobID=10 has a link that points back to itself	RESOLVED
	Log how many new domains are created during a domain-ingest	RESOLVED
	When starting to generate jobs for a snapshotharvest set numevents=1 to avoid a restart of the process	CLOSED
	Definitions-snapshot-harvests.jsp need to be sorted by harvestname	RESOLVED
	Compression is not currently a choice when building our metadata files	RESOLVED
	Use 24 hour clock everywhere in English GUI	CLOSED
	Merge new CSS in	RESOLVED
	the SeedUriDomainnameQueueAssignmentPolicy ignores source tag when handling dns urls	RESOLVED
	Implement NASFetchDNS to allow to the use of local host overrides	RESOLVED
	harvestInfo.performer in warcinfo records is not included in harvestInfo.xml	RESOLVED
	Takes 5 minutes before link to H3 server appears in Running Jobs	RESOLVED
	Garbage content on job-details-page	RESOLVED
	GUI now needs a default context	CLOSED

30 issues

Known issues

T	Key	P	Summary	Fix Version/s
	NAS-2582		DeduplicateToCDXAdapter fails to identify new dedup format	5.2.1
	NAS-2584		Deduplication via settings doesn't work	5.2.2
	NAS-2640		The caching functionality of the H3 remote access does not scale	5.4
	NAS-2491		QUICKSTART db not up to date	5.1
	NAS-2587		Software stated in the metadata files warcinfo records cannot be easily parsed	5.3
	NAS-2585		NetarchiveResourceStore doesn't handle revisits well	5.3
	NAS-2641		In H3 Remote Access, the Next/Previous regex search in crawllogs doesn't work	5.4
	NAS-2596		H3 pauseAtStart bean property ignored by Harvest Controller	5.3
	NAS-2496		Redirect for jp.dk fails in test wayback	5.3
	NAS-2644		The seedslists in NAS are sorted alfabetically which mess with the correct crawlorder of some pages where harvesting of the login-page needs to be crawled first	5.5
	NAS-2583		Wayback Indexed files don't get an IndexedDate	5.5
	NAS-2578		Inconsistent SystemTest result	5.5
	NAS-2643		When H3 remote access is no longer possible, don't show a status 500 error page	5.4.3
	NAS-2577		NetarchiveCacheResourceStore missing dependency	

14 issues