NetarchiveSuite 5.4.x Release Notes

Most-recent updates for 5.4.1:

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

5.4.2 Release Date: 2018-06-15

BugFix Release 5.4.2

This release addresses issue **NAS-2514** - Getting issue details... STATUS which resulted in many url's receiving crawl-status code -50 in some harvests. It is only relevant for users of SeedUriDomainnameQueueAssignmentPolicy. The fix is in two parts:

- A new QuotaEnforcer implementation dk.netarkivet.harvester.harvesting.PrerequisiteIgnoringQuotaEnforcer which can be used in a crawler-bean harvest-template, and which never enforces harvesting quotas on prerequisite url's (typically dns lookups and robots.txt), and
- An alteration to SeedUriDomainnameQueueAssignmentPolicy to ensure that dns queries are queued on the same queue as other url's for the same seed. This appears to work around an undocumented race condition in heritrix which was causing many crawl failures.

BugFix Release 5.4.1

NAS 5.4.1 is a Bug-Fix release addressing some issues found during the Acceptance Test phase of NAS 5.4. The issues addressed are

- A memory leak introduced by a new feature in NAS 5.4 (**NAS-2614** - Getting issue details... STATUS) to manage the number of jobs on the JMS queues, and
- An error in the functionality for searching/browsing in the frontier of running jobs
- Introduction of a new setting (`settings.harvester.indexserver.tryToMigrateDuplicationRecords`), a switch, to disable new functionality associated with the Danish netarchive's project to compress their archive. This functionality caused an unnecessary slowdown in indexing functionality, but is now disabled by default.

The functionality for browsing in the Heritrix frontier is still somewhat experimental and is in need of a usability overhaul. This is a priority for a future release.

Key	Summary	Status
NAS-2754	Problem listing/deleting from frontier	CLOSED
NAS-2751	The HarvestStatusReceiver.getCount method has a memoryLeak	CLOSED
NAS-2752	We always look for duplicationmigration records during indexing	CLOSED

3 issues

Highlights in 5.4

- NetarchiveSuite now ships with a customised version of Heritrix 3, forked from the version maintained by Kristinn Sigurdsson at the National Library of Iceland.
- The integration between the NetarchiveSuite Web interface and Heritrix 3 has been much improved, both in regard to scaling and usability.
- There is significant improvement to the job generation algorithm, so that the production of spurious duplicate jobs is now largely eliminated.
- Support for Heritrix1 has now been removed from the distribution.
- You can now define a limit to how many jobs are submitted to each jobchannel

simultaneously, if you enable limitSubmittedJobsInQueue by setting *settings.harvester.scheduler.limitSubmittedJobsInQueue* to true. The default value if you enable this is one job at a time. You can change this value by overriding the *settings.harvester.scheduler.submittedJobsInQueueLimit*. The latter setting is ignored, if *limitSubmittedJobsInQueue* is false, which is the default setting.

- The setting *settings.harvester.scheduler.jobgenerationperiode* has been renamed *settings.harvester.scheduler.jobgenerationperiod* (default value is still 60 a.k.a 1 minute)
- Added new setting to choose between filtering methods on History/Harveststatus-running.jsp: *settings.webinterface.runningjobsFilteringMethod (*default*: database* alternative*: cachedLogs)*

Upgrading from previous releases of Netarchivesuite

- Upgrading the database: After finishing the installation of NetarchiveSuite and starting it for the first time, please go the server where GUIApplication and HarvestJobManager is installed and run:

  ```
  cd NAS_INSTALLDIR/conf
  bash update_external_harvest_database.sh
  ```

  Please examine the INSTALLDIR/update_external_harvest_database.log for any errors.

Issues resolved in release 5.4

Key	Summary	Status
NAS-2682	Multiple duplicate Jobs Created	CLOSED
NAS-2694	Errors Starting GUI - some SiteSections missing	CLOSED
NAS-2726	The methods in dk.netarkivet.viewerproxy.webinterface.Reporting does not support metadata files using the BNF naming	RESOLVED
NAS-2731	History/Harveststatus-running.jsp shows malformed webform when using French as GUI languages	RESOLVED
NAS-2732	Columns Bytes Harvested / Documents Harvested / Stopped due to on page Details for job XXX are always empty after the job termination.	CLOSED
NAS-2733	Show only jobs harvesting domain functionality doesn't work - can't find cached crawl.log to search in	RESOLVED
NAS-2640	The caching functionality of the H3 remote access does not scale	CLOSED
NAS-2693	NetarchiveResourceStore does not set "Origin" Metadata	CLOSED
NAS-2702	The jobgeneration of Snapshot harvests using the DefaultJobGenerator.java generates too many jobs	CLOSED
NAS-1527	Thousand separators requested in user interface	CLOSED
NAS-2463	Find which jobs are harvesting a given domain	RESOLVED
NAS-2534	Remove unwanted folders on job restart	CLOSED
NAS-2560	NPE during closing down of the toethread	CLOSED
NAS-2614	Multiple jobs submitted simultaneously	CLOSED
NAS-2641	In H3 Remote Access, the Next/Previous regex search in crawllogs doesn't work	CLOSED
NAS-2642	The Frontier search in H3 remote access must be paged instead of showing it all - which crashes the browser	CLOSED
NAS-2649	harvestInfo.XXXX fields are not added in warcinfo records for resubmitted jobs	RESOLVED
NAS-2674	During the harvester Registration phase the harvester shuts down if its channel ID is unknown	CLOSED

NAS-2409	A shutdown.txt in a harvestfolder does not prevent a restart, when the job fails	CLOSED
NAS-2579	Much "Invalid tld" logspam	CLOSED
NAS-2638	Separate the H3 remote access module from the GUI application	READY FOR REVIEW
NAS-2647	H3 remote access code does not respect user language	CLOSED
NAS-2648	Indexserver fails during indexing to substitute properly duplicate-records in old metadata files with the old duplicate annotation	RESOLVED
NAS-2650	harvestInfo.origHarvestDefinitionComments are missing in warcinfo records	RESOLVED
NAS-2662	Progression/Queues in H3 Remote Access	RESOLVED
NAS-2676	The regexp used by Reporting.getMetadataCDXRecordsForJob(jobID) in NetarchiveSuite is wrong	CLOSED
NAS-2677	Remove H1 support from Netarchivesuite	RESOLVED
NAS-2678	NasWARCProcessor throws ugly NullPointerException if harvestInfo missing from template	CLOSED
NAS-2681	Improve the Postgresql documentation	RESOLVED
NAS-2686	Alway create a metadata-warcfile even if Heritrix3 doesn't create any (w)arc files	CLOSED
NAS-2687	Incomplete lines in the duplicationMigration are not caught in RawMetadataCache.migrateDuplicates()	CLOSED
NAS-2688	Trim Domainnames before ingest	RESOLVED
NAS-2690	The function "Browse only relevant crawl-log lines for this domain" is faulty	CLOSED
NAS-2692	Create dummyHarvestControllerApplication for scale-testing	RESOLVED
NAS-2696	Update copyright statement in headers to "Copyright (C) 2005 - 2018"	CLOSED
NAS-2701	StreamUtils.getInputStreamAsString() and StreamUtils.copyInputStreamToJspWriter() throw IndexOutOfBounds exceptions	CLOSED
NAS-2704	Typo and wrong default value for setting settings.harvester.scheduler.jobgenerationperiode	CLOSED
NAS-2710	Freespace Provider for ONB needs	RESOLVED
NAS-2711	Improve usability on pages Definitions-edit-snapshot-harvest.jsp, Definitions-selective-harvests.jsp, and Definitions-snapshot-harvests.jsp	CLOSED
NAS-2723	The links in https://sbforge.org/display/NASDOC/Installation+of+the+Quickstart+system to releases are wrong	RESOLVED
NAS-2724	Quickstart setup cannot browse reports jobreports and crawllog-lines	RESOLVED
NAS-2728	Message "Branch in settings not found" shown during deploy	RESOLVED
NAS-2729	Stopping deduplication by setting doesn't work as expected ...	CLOSED
NAS-2734	SeedUriDomainnameQueueAssignmentPolicy.getKeyFromSeed() throws org.apache.commons.httpclient.URIException during standard test harvest of netarkivet.dk	RESOLVED
NAS-2735	When searching for jobs harvesting a specific domain, it resets the value of the text field after the search	RESOLVED
NAS-2737	Not possible to extract domainname from URL: www.netarkivet.dk	RESOLVED
NAS-2741	WARN d.n.harvester.datamodel.Domain.setCrawlerTraps - 0 errors were found:	CLOSED
NAS-2744	H3 version in dk.netarkivet.common.Constants#getHeritrix3VersionString() returns wrong version	RESOLVED

48 issues