Heritrix Release Notes

Internet Archive

Table of Contents

1. Release 1.14.4 - 2010-04-30

Release 1.14.4 is a bugfix release with 44 tracked fixes and small improvements.

Other details on this release, including a list of all tracked issues addressed, are available at: Release Notes - Heritrix 1.14.4 at project wiki [https://webarchive.jira.com/wiki/display/Heritrix/Release+Notes+-+1.14.4]

2. Release 1.14.3 - 2009-03-02

Release 1.14.3 is a bugfix release with 14 tracked fixes and small improvements.

Other details on this release, including a list of all tracked issues addressed, are available at: Release Notes - Heritrix 1.14.3 at project wiki [http://webteam.archive.org/confluence/display/Heritrix/Release+Notes+-+1.14.3]

3. Release 1.14.2 - 2008-08-06

Release 1.14.2 is a bugfix release with 6 tracked fixes.

Other details on this release, including a list of all tracked issues addressed, are available at: Release Notes - Heritrix 1.14.2 at project wiki [http://webteam.archive.org/confluence/display/Heritrix/Release+Notes+-+1.14.2]

4. Release 1.14.1 - 2008-08-06

Release 1.14.1 is a bugfix release with 25 tracked fixes and small requested/contributed features.

Other details on this release, including a list of all tracked issues addressed, are available at: Release Notes - Heritrix 1.14.1 at project wiki [http://webteam.archive.org/confluence/display/Heritrix/Release+Notes+-+1.14.1]

5. Release 1.14.0 - 2008-04-27

Release 1.14.0 adds a number of small features to the Heritrix 1.x line, most notably upgrading support for the WARC archived-web-content format to version 0.17 (ISO Committee Draft). This release also includes 41 bug fixes or other incremental improvements.

Other details on this release, including a list of all tracked issues addressed, are available at: Release Notes - Heritrix 1.14.0 at project wiki [http://webteam.archive.org/confluence/display/Heritrix/Release+Notes+-+1.14.0]

6. Release 1.12.1 - 2007-05-06

Release 1.12.1 is primarily a bug-fix release. A total of 21 bugs have been addressed, including an object-retention issue when running multiple jobs in sequence, a failure to enforce fetch timeouts on certain malformed server responses, and decide-rule overrides disappearing when basing jobs on earlier jobs/profiles. There are also several enhancements, such as reporting the volume of data evaluated but not transferred/stored when using duplicate-reduction features.

6.1. Contributors

Aside from the usual suspects [http://crawler.archive.org/team-list.html], the following people contributed to this release:

- Ahmed Ghouzia

6.2. All tracked changes

A dynamic list of all tracked changes marked as fixed in 1.12.1 is available at: Issues with 'Fix Version' 1.12.1.
[http://webteam.archive.org/confluence/display/Heritrix/Issues+with+%27Fix+Version%27+1.12.1]

7. Release 1.12.0 - 2007-03-16

Release 1.12.0 is the first of several planned releases enhancing Heritrix with "smart crawler" functionality. In this release, the theme has been offering new options to reduce the amount of duplicate content crawled and stored when recrawling sites at regular intervals. A number of other enhancements and bug fixes are also included.

7.1. Contributors

Aside from the usual suspects [http://crawler.archive.org/team-list.html], the following contributed to this release:

- Oskar Grenholm
- Doug Judd

7.2. Notes

With this release, Heritrix project issue-tracking has moved from Sourceforge to a JIRA-based system at http://webteam.archive.org/jira/browse/HER [http://webteam.archive.org/jira/browse/HER].

Those using Heritrix in a Hadoop environment may be interested in Doug Judd's HDFSWriterProcessor [http://www.zvents.com/labs/hdfs_writer_processor], for storing crawled content directly into HDFS, the Hadoop Distributed FileSystem [http://lucene.apache.org/hadoop/].

7.3. Known Limitations/Issues

7.3.1. java.io.IOException: No locks available

See Section 11.1.1, “java.io.IOException: No locks available” in 1.8.0 Release Notes.

7.3.2. Older Checkpoints

Checkpoints from earlier versions are generally not supported for resume in later versions.

7.3.3. Older configurations (order.xml, etc.)

Crawler configuration files from jobs in previous versions may work in 1.12.0, though missing new settings will be set to their default values, and obsolete old settings will generate log warnings. Re-creating configurations from defaults or hand-editting to match newer files is recommended.

7.4. Changes

7.4.1. Duplication reduction features

A collection of Processors, including the FetchHistoryProcessor, PersistProcessor, and its subclasses, may be used together with new options on the FetchHTTP and writer processors to carry information forward between crawls and collect less duplicate content on later recrawls. The project wiki features notes on using the new duplication-reduction functionality [http://webteam.archive.org/confluence/display/Heritrix/Feature+Notes+-+1.12.0].

7.4.2. DecideRules have replaced Filters on Processors

All Processors which used internal Filters for differentially acting on URIs now use DecideRules instead. In those cases where a DecideRule replacement for a Filter is not yet available, a legacy Filter can be wrapped in a FilterDecideRule to preserve prior functionality. In a future release, all Filters will be removed in favor of equivalent DecideRules.

7.4.3. WARC

ExperimentalWARCWriter has been updated to match proposed WARC version "WARC/0.12" (revision H1.12-RC1) [http://archive-access.svn.sourceforge.net/viewvc/*checkout*/archive-access/branches/gjm_warc_0_12/warc/warc_file_format.html]. The implementation as of Heritrix 1.10.x remains for reference as org.archive.io.warc.v10.ExperimentalV10WARCWriterProcessor. The WARC format remains under discussion.

7.4.4. Kw3WriterProcessor

Oskar Grenholm of the Swedish National Library has contributed a module that writes the results of successful fetches to files on disk. These files are MIME-files of the type used by the Swedish National Library's Kulturarw3 web harvesting [http://www.kb.se/kw3/].

7.4.5. All tracked changes

A dynamic list of all tracked changes marked as fixed in 1.12.0 is available at: Issues with 'Fix Version' 1.12.0.
[http://webteam.archive.org/confluence/display/Heritrix/Issues+with+%27Fix+Version%27+1.12.0]

8. Release 1.10.2 - 2007-01-15

This is primarily a bug-fix release, with a couple of new features, provided before a number of significant changes to the Heritrix project that will require developer and crawl operator adjustments. Post-

1.10.2, Heritrix source code control, issue tracking, and build process will migrate to new systems. Also, updates to core classes, especially with regard to the settings architecture, will noticeably break backward compatibility with 1.10.2 and prior crawler settings files and formats.

8.1. Contributors

- Olaf Freyer
- Max Schöfmann

8.2. Changes

8.2.1. Jericho HTML Extractor

Olaf Freyer has contributed an HTML Extractor named JerichoExtractorHTML based on the Jericho HTML Parser. Following is a quote from the JerichoExtractorHTML class comment describing how the new Extractor differs from ExtractorHTML, its advantages and downsides: " This extractor extends ExtractorHTML and mimics its workflow - but has some substantial differences when it comes to internal implementation. Instead of heavily relying upon java regular expressions it uses a real html parser library - namely Jericho HTML Parser (http://jerichohtml.sourceforge.net). Using this parser it can better handle broken html (i.e. missing quotes) and also offer improved extraction of HTML form URLs (not only extract the action of a form, but also its default values). Unfortunately this parser also has one major drawback - it has to read the whole document into memory for parsing, thus has an inherent OOME risk. This OOME risk can be reduced/eleminated by limiting the size of documents to be parsed (i.e. using NotExceedsDocumentLengthTresholdDecideRule). Also note that this extractor seems to have a lower overall memory consumption compared to ExtractorHTML. (still to be confirmed on a larger scale crawl) "

Table 1. All Tracked Changes

ID	Type	Summary	Open Date	By	Filer
913002 [https://sourceforge.net/tracker/index.php?func=detail&aid=913002&group_id=73833&atid=539102]	Add	Make ExtractorHTML aggressiveness configurable	2004-03-09	gojomo	gojomo
1573708 [https://sourceforge.net/tracker/index.php?func=detail&aid=1573708&group_id=73833&atid=539102]	Add	[Contrib] JerichoExtractorHTML	2006-10-09	nobody	pandae
1633458	Add	[arcreader] Sup-	2007-01-11	stack	stack

ID	Type	Summary	Open Date	By	Filer
[https://sourceforge.net/tracker/index.php?func=detail&aid=1573708&group_id=73833&atid=539102]		port for s3 and streaming improvements			
1629242 [https://sourceforge.net/tracker/index.php?func=detail&aid=1629242&group_id=73833&atid=539099]	Fix	filehandle leak: ReplayInputStream/ BufferedSeekInputStream	2007-01-05	karl-ia	gojomo
1218961 [https://sourceforge.net/tracker/index.php?func=detail&aid=1218961&group_id=73833&atid=539099]	Fix	"failed get of replay" in ExtractorHTML... usu: UTF-16BE	2005-06-11	karl-ia	gojomo
996161 [https://sourceforge.net/tracker/index.php?func=detail&aid=996161&group_id=73833&atid=539099]	Fix	Fix DNSJava issues (memory)	2004-07-22	karl-ia	gojomo
1477371 [https://sourceforge.net/tracker/index.php?func=detail&aid=1477371&group_id=73833&atid=539099]	Fix	ExtractorDOC wants whole doc in memory	2006-04-26	paul_jack	gojomo
1618928 [https://sourceforge.net/tracker/index.php?func=detail&aid=1618928&group_id=73833&atid=539099]	Fix	Do not allow http:/ and https:/ urls	2006-12-19	stack-sf	stack-sf
1596176 [https://sourcefo	Fix	NotMatchesListRegExpDeci-	2006-11-14	nobody	pandae

ID	Type	Summary	Open Date	By	Filer
rge.net/tracker/index.php?func=detail&aid=1596176&group_id=73833&atid=539099]		deRule extends wrong class			
1593540 [https://sourceforge.net/tracker/index.php?func=detail&aid=1593540&group_id=73833&atid=539099]	Fix	NPE in quotaEnforcer.checkQuotas	2006-11-09	nobody	svc
1587413 [https://sourceforge.net/tracker/index.php?func=detail&aid=1587413&group_id=73833&atid=539099]	Fix	[PATCH] Webapp doesn't find profiles and ignores jobsdir	2006-10-30	nobody	nobody
1572391 [https://sourceforge.net/tracker/index.php?func=detail&aid=1572391&group_id=73833&atid=539099]	Fix	SURTs for IP-address URIs unhelpful	2006-10-06	gojomo	gojomo
1501810 [https://sourceforge.net/tracker/index.php?func=detail&aid=1501810&group_id=73833&atid=539099]	Fix	NPE in FetchHTTP.saveCookies	2006-06-06	gojomo	stack-sf
1633117 [https://sourceforge.net/tracker/index.php?func=detail&aid=1501810&group_id=73833&atid=539099]	Fix	Useragent compare because of case in RobotsExclusionPolicy	2007-01-11	stack-sf	stack-sf

9. Release 1.10.1 - 2006-09-27

Bug fixes.

9.1. Changes

Table 2. All Tracked Changes

ID	Type	Summary	Open Date	By	Filer
1566041 [http://sourceforge.net/tracker/index.php?func=detail&aid=1566041&group_id=73833&atid=539099]	Fix	In 1.10.0, ARC serialno starts at 1 rather than 0	2006-09-26	stack-sf	stack-sf
1558157 [http://sourceforge.net/tracker/index.php?func=detail&aid=1558157&group_id=73833&atid=539099]	Fix	JNDI registration fails in 1.10.0	2006-09-13	stack-sf	stack-sf
1560782 [http://sourceforge.net/tracker/index.php?func=detail&aid=1560782&group_id=73833&atid=539099]	Fix	StripWWWN-Rule does not work as advertised	2006-09-18	stack-sf	stack-sf
1563592 [http://sourceforge.net/tracker/index.php?func=detail&aid=1563592&group_id=73833&atid=539099]	Fix	ARCWriter constructor had its public access removed	2006-09-25	stack-sf	stack-sf

10. Release 1.10.0 - 2006-09-11

Release 1.10.0 adds new configuration options, experimental new protocol and format support, and lots of fixes. 43 tracked bugs have been fixed and 35 feature requests added.

Release 1.10.0 requires JDK 1.5.x ("Java 5") Java facilities.

10.1. Contributors

Aside from the usual suspects [http://crawler.archive.org/team-list.html], the following contributed to this release:

- Eric C Jensen
- Olaf Freyer
- Karl Wright (of MetaCarta)
- Frank McCown (of Old Dominion University)
- Max Schöfmann
- Søren Vejrup Carlsen (of Royal Library, Denmark)

10.2. Known Limitations/Issues

10.2.1. java.io.IOException: No locks available

See Section 11.1.1, “java.io.IOException: No locks available” in 1.8.0 Release Notes.

10.3. Pre-1.10.0 checkpoints

For sure 1.8.0 checkpoints will not be recoverable with 1.10.0.

10.4. Changes

10.4.1. No default login/password for web UI and JMX

The old default login of 'admin' and password of 'letmein' for access to the crawler web UI (and JMX agent control) have been eliminated. It is now necessary to specify an access username and password to start Heritrix. This may be done with the -a or --admin command-line argument or via the system property 'heritrix.cmdline.admin'. (These each take a colon-separated username and password, like 'username:password'.)

10.4.2. Web UI binds to localhost only by default

Previously, the Jetty web server that runs the Heritrix web UI listened on all available network interfaces. In 1.10.0, Jetty will only bind to localhost by default. The -b or --bind command-line argument can be used to specify a different interface or list of interfaces to bind to instead. You may specify "-b /" to get the old behavior -- binding on all interfaces -- but only take this step after reading section 2.3 of the User Manual, "Security Considerations".

10.4.3. QuotaEnforcer 'force-retire' option

The optional QuotaEnforcer processor has a new setting, 'force-retire', which is by default 'true', and changes the default behavior of QuotaEnforcer. Previously, when a URI was noted as being over-quota, it would be marked with a special over-quota failure code which caused it to complete processing as an error. As a result, all over-quota URIs would quickly be finished as errors and appear in the crawl.log, but there would be no opportunity to raise the quota and continue crawling.

The new default behavior instead marks the URI with a directive requesting its frontier queue be retired. If the frontier supports this directive, the URI will be returned to its queue as if never tried, and the

whole queue retired from active crawling. This offers the opportunity to raise the quota and continue crawling the URI and others of its queue. (All settings changes cause all retired queues to be reevaluated.) However, the over-quota URIs will not appear as errors in the crawl.log.

If the old behavior is preferred, set 'force-retire' to 'false'.

10.4.4. URL canonicalization changes

In 1.10.0, URL canonicalization has changed in two ways. First, the stripping of sessionids has improved [See Stripping sessionid can leave behind doubled ampersands [http://sourceforge.net/tracker/index.php?func=detail&aid=1550797&group_id=73833&atid=539099]]. Previous, if the sessionid was in the middle of a query string bookended by other query parameters, canonicalization would leave behind the encasing ampersands: E.g. If the URL `http://a.com/?a=1&sid=00000000000000000000000000000000&b=1` was passed through canonicalization, the result would be: `http://a.com/?a=1&&b1`. This has been fixed so that the result will now be: `http://a.com/?a=1&b1`.

The second change, [1550805] Add stripping of coldfusion sessionids [http://sourceforge.net/tracker/index.php?func=detail&aid=1550805&group_id=73833&atid=539102], adds the new coldfusion sessionid stripper to the list of default canonicalization rules.

We bring your attention to these seemingly minor changes because for those of you running regular crawls, with both of the above changes in place, depending on the type of crawl, there should be a reduction in overall the number of (duplicate) pages crawled.

10.4.5. WARC

This release includes experimental WARC readers and writers. Be warned that both code and specification are not yet final and so are both subject to change with no guarantees of backward compatibility: i.e. newer readers may not be able to read WARCs written with older writers. See the org.archive.io.warc [/apidocs/org/archive/io/warc/package-summary.html] package documentation for more on the current state of code including documentation of initial version of `Arc2Warc` and `Warc2Arc` tools.

10.4.6. FTP

This release also include experimental support for FTP. This support is disabled by the default heritrix configuration. See the User Guide for information on how to enable FTP.

Table 3. All Tracked Changes

ID	Type	Summary	Open Date	By	Filer
1545462 [http://sourceforge.net/tracker/index.php?func=detail&aid=1545462&group_id=73833&atid=539102]	Add	Experimental WARC Readers and Writers	2006-08-23	stack-sf	stack-sf
1494491 [http://sourceforge.net/tracker/index.php?func=detail&aid=1494	Add	path/ role-sensitive robots (eg ignore for inline images/css)	2006-05-24	karl-ia	gojomo

ID	Type	Summary	Open Date	By	Filer
491&group_id=73833&atid=539102]					
1550849 [http://sourceforge.net/tracker/index.php?func=detail&aid=1550849&group_id=73833&atid=539102]	Add	'Implied' URI extractor (eg, YouTube)	2006-09-01	karl-ia	gojomo
1549665 [http://sourceforge.net/tracker/index.php?func=detail&aid=1549665&group_id=73833&atid=539102]	Add	Add experimental Warc2Arc and Arc2Warc scripts	2006-08-30	stack-sf	stack-sf
1546829 [http://sourceforge.net/tracker/index.php?func=detail&aid=1546829&group_id=73833&atid=539102]	Add	Secure admin UI: Bind cmd-line argument	2006-08-25	karl-ia	stack-sf
1545600 [http://sourceforge.net/tracker/index.php?func=detail&aid=1545600&group_id=73833&atid=539102]	Add	remove default admin user-name/password	2006-08-23	karl-ia	gojomo
1536441 [http://sourceforge.net/tracker/index.php?func=detail&aid=1536441&group_id=73833&atid=539102]	Add	hash-based CrawlMapper	2006-08-08	karl-ia	gojomo
1535744 [http://sourceforge.net/tracker/index.php?func=detail&aid=1535744&group_id=73833&atid=539102]	Add	force reread of disk settings (for out-of-JVM/bulk changes)	2006-08-06	karl-ia	gojomo
1534280 [http://sourceforge.net/tracker/in	Add	scriptable (beanshell) Processor, Deci-	2006-08-03	gojomo	gojomo

ID	Type	Summary	Open Date	By	Filer
dex.php?func=detail&aid=1534280&group_id=73833&atid=539102]		deRule options			
1522112 [http://sourceforge.net/tracker/index.php?func=detail&aid=1522112&group_id=73833&atid=539102]	Add	CrawlMapper skip mapping 'E'mbeds (etc)	2006-07-13	karl-ia	gojomo
1520269 [http://sourceforge.net/tracker/index.php?func=detail&aid=1520269&group_id=73833&atid=539102]	Add	keep over-limit (-500X) URIs in queues (don't 'finish/log)	2006-07-10	karl-ia	gojomo
1387423 [http://sourceforge.net/tracker/index.php?func=detail&aid=1387423&group_id=73833&atid=539102]	Add	[arcreader] Fetch records and iterate remote ARCs	2005-12-21	stack-sf	stack-sf
1351778 [http://sourceforge.net/tracker/index.php?func=detail&aid=1351778&group_id=73833&atid=539102]	Add	favicon.ico for heritrix web ui	2005-11-08	gojomo	gojomo
1209724 [http://sourceforge.net/tracker/index.php?func=detail&aid=1209724&group_id=73833&atid=539102]	Add	[contrib] Add BigMapFactory.getSynchronizedBigMap	2005-05-27	gojomo	ck-heritrix
1526781 [http://sourceforge.net/tracker/index.php?func=detail&aid=1526781&group_id=73833&atid=539102]	Add	broader rotation / wider 'front-line' frontier queue option	2006-07-21	karl-ia	gojomo
1092496	Add	crawl.log	2004-12-28	stack-sf	gojomo

ID	Type	Summary	Open Date	By	Filer
[http://sourceforge.net/tracker/index.php?func=detail&aid=1092496&group_id=73833&atid=539102]		should have hash of DNS records			
1006194 [http://sourceforge.net/tracker/index.php?func=detail&aid=1006194&group_id=73833&atid=539102]	Add	FTP fetching	2004-08-09	karl-ia	gojomo
1550805 [http://sourceforge.net/tracker/index.php?func=detail&aid=1550805&group_id=73833&atid=539102]	Add	Add stripping of coldfusion sessionids -- add to default lis	2006-09-01	stack-sf	stack-sf
1547390 [http://sourceforge.net/tracker/index.php?func=detail&aid=1547390&group_id=73833&atid=539102]	Add	[contrib] patch to allow setting local IP to bind fetch from	2006-08-26	stack-sf	ecjensen
1545847 [http://sourceforge.net/tracker/index.php?func=detail&aid=1545847&group_id=73833&atid=539102]	Add	[contrib] allow to specify alternative conf location	2006-08-24	stack-sf	pandae
1545840 [http://sourceforge.net/tracker/index.php?func=detail&aid=1545840&group_id=73833&atid=539102]	Add	[contrib] ContentLengthFilter	2006-08-24	stack-sf	pandae
1537507 [http://sourceforge.net/tracker/index.php?func=detail&aid=1537507&group_id=73833&atid=539102]	Add	Add checkpointing selftest	2006-08-09	stack-sf	stack-sf

ID	Type	Summary	Open Date	By	Filer
1535116 [http://sourceforge.net/tracker/index.php?func=detail&aid=1535116&group_id=73833&atid=539102]	Add	Add creation/deletion of Heritrix instances to UI	2006-08-05	karl-ia	stack-sf
1530557 [http://sourceforge.net/tracker/index.php?func=detail&aid=1530557&group_id=73833&atid=539102]	Add	[contrib] Enhanced UI seed and crawl reports	2006-07-28	stack-sf	stack-sf
1523276 [http://sourceforge.net/tracker/index.php?func=detail&aid=1523276&group_id=73833&atid=539102]	Add	Should support depth-first search priority scheduling (patch	2006-07-15	stack-sf	ecjensen
1518583 [http://sourceforge.net/tracker/index.php?func=detail&aid=1518583&group_id=73833&atid=539102]	Add	Improved handling when alloted runtime is exceeded	2006-07-07	kristinn_sig	kristinn_sig
1514538 [http://sourceforge.net/tracker/index.php?func=detail&aid=1514538&group_id=73833&atid=539102]	Add	(contrib) Provide Windows batch file version of scripts	2006-06-29	nobody	ecjensen
1510807 [http://sourceforge.net/tracker/index.php?func=detail&aid=1510807&group_id=73833&atid=539102]	Add	[contrib] Have Heritrix UI bind to localhost only	2006-06-22	karl-ia	stack-sf
1505111 [http://sourceforge.net/tracker/index.php?func=detail&aid=1505111&group_id=73833&atid=53	Add	Make deciding-default profile the default profile	2006-06-12	stack-sf	stack-sf

ID	Type	Summary	Open Date	By	Filer
9102]					
1489231 [http://sourceforge.net/tracker/index.php?func=detail&aid=1489231&group_id=73833&atid=539102]	Add	Move to java 5.0/1.5.0	2006-05-15	nobody	stack-sf
1388295 [http://sourceforge.net/tracker/index.php?func=detail&aid=1388295&group_id=73833&atid=539102]	Add	[contrib] Throttling on a per-document basis	2005-12-22	karl-ia	stack-sf
1153882 [http://sourceforge.net/tracker/index.php?func=detail&aid=1153882&group_id=73833&atid=539102]	Add	change username/password after launch	2005-02-28	karl-ia	gojomo
1058324 [http://sourceforge.net/tracker/index.php?func=detail&aid=1058324&group_id=73833&atid=539102]	Add	Show old crawl reports in UI (Was: Reports on finished...)	2004-11-01	nobody	stack-sf
986985 [http://sourceforge.net/tracker/index.php?func=detail&aid=986985&group_id=73833&atid=539102]	Add	Fix API to allow ARCWriter replacement	2004-07-07	stack-sf	stack-sf
1540381 [http://sourceforge.net/tracker/index.php?func=detail&aid=1540381&group_id=73833&atid=539099]	Fix	proxying of https gives errors/garbage/later problems	2006-08-14	karl-ia	gojomo
1534082 [http://sourceforge.net/tracker/index.php?func=detail&aid=1534	Fix	override of user-agents and masquerade not working	2006-08-03	karl-ia	ia_igor

ID	Type	Summary	Open Date	By	Filer
082&group_id=73833&atid=539099]					
1495253 [http://sourceforge.net/tracker/index.php?func=detail&aid=1495253&group_id=73833&atid=539099]	Fix	multiple usage of same arc id number within same crawl	2006-05-25	karl-ia	ia_igor
1533571 [http://sourceforge.net/tracker/index.php?func=detail&aid=1533571&group_id=73833&atid=539099]	Fix	Checkpointing is broken (Parts 1 and 2)	2006-08-02	stack-sf	stack-sf
1511596 [http://sourceforge.net/tracker/index.php?func=detail&aid=1511596&group_id=73833&atid=539099]	Fix	incorrect resolving relative links from flash files (swf)	2006-06-23	karl-ia	ia_igor
1510289 [http://sourceforge.net/tracker/index.php?func=detail&aid=1510289&group_id=73833&atid=539099]	Fix	CSS keywords are case sensetive in extraction	2006-06-21	gojomo	cathcart
1489132 [http://sourceforge.net/tracker/index.php?func=detail&aid=1489132&group_id=73833&atid=539099]	Fix	Contain HttpClient HttpParser's OutOfMemoryError risk	2006-05-15	karl-ia	gojomo
1442679 [http://sourceforge.net/tracker/index.php?func=detail&aid=1442679&group_id=73833&atid=539099]	Fix	HTMLExtractor and application/xhtml+xml type?	2006-03-03	karl-ia	gojomo
1549627 [http://sourceforge.net/tracker/in	Fix	Archive file serialnumber is always 1 after	2006-08-30	stack-sf	stack-sf

ID	Type	Summary	Open Date	By	Filer
dex.php?func=detail&aid=1549627&group_id=73833&atid=539099]		checkpoint			
1546808 [http://sourceforge.net/tracker/index.php?func=detail&aid=1546808&group_id=73833&atid=539099]	Fix	Don't resume crawl after checkpoint if state is 'pausing'	2006-08-25	stack-sf	stack-sf
1542933 [http://sourceforge.net/tracker/index.php?func=detail&aid=1542933&group_id=73833&atid=539099]	Fix	adjust prominence of instance/identifier info/tab	2006-08-18	karl-ia	gojomo
1540030 [http://sourceforge.net/tracker/index.php?func=detail&aid=1540030&group_id=73833&atid=539099]	Fix	FetchDNS IOException: Stream closed	2006-08-14	stack-sf	stack-sf
1538489 [http://sourceforge.net/tracker/index.php?func=detail&aid=1538489&group_id=73833&atid=539099]	Fix	HeritrixProtocolSocketFactory synchronization causes delays	2006-08-11	stack-sf	gojomo
1534153 [http://sourceforge.net/tracker/index.php?func=detail&aid=1534153&group_id=73833&atid=539099]	Fix	don't insist on robots.txt if it need not be honored	2006-08-03	karl-ia	gojomo
1532787 [http://sourceforge.net/tracker/index.php?func=detail&aid=1532787&group_id=73833&atid=539099]	Fix	OnDomainsDecideRule not working as expected	2006-08-01	gojomo	gojomo
1532665	Fix	AddRedirect-	2006-08-01	gojomo	gojomo

ID	Type	Summary	Open Date	By	Filer
[http://sourceforge.net/tracker/index.php?func=detail&aid=1532665&group_id=73833&atid=539099]		FromRoot-ServerToScope not working as expected			
1519056 [http://sourceforge.net/tracker/index.php?func=detail&aid=1519056&group_id=73833&atid=539099]	Fix	IPQueueAs-signmentPolicy broken by method signature mismatch	2006-07-07	gojomo	gojomo
1514716 [http://sourceforge.net/tracker/index.php?func=detail&aid=1514716&group_id=73833&atid=539099]	Fix	heritrix fails to save accept-headers in an override	2006-06-29	karl-ia	magin-ia
1511624 [http://sourceforge.net/tracker/index.php?func=detail&aid=1511624&group_id=73833&atid=539099]	Fix	NoOnDomains-DecideRule/ NotOnHostsDecideRule superclass wrong	2006-06-23	karl-ia	gojomo
1482210 [http://sourceforge.net/tracker/index.php?func=detail&aid=1482210&group_id=73833&atid=539099]	Fix	CachedBdbMap.keySet inefficient or broken	2006-05-04	karl-ia	gojomo
1475798 [http://sourceforge.net/tracker/index.php?func=detail&aid=1475798&group_id=73833&atid=539099]	Fix	ARCReader#read(byte [], off, len) broke for non-null offset	2006-04-24	stack-sf	stack-sf
1189825 [http://sourceforge.net/tracker/index.php?func=detail&aid=1189825&group_id=73833&atid=539099]	Fix	ARC problem causing .invalid suffix needs better reporting	2005-04-25	paul_jack	gojomo

ID	Type	Summary	Open Date	By	Filer
1056919 [http://sourceforge.net/tracker/index.php?func=detail&aid=1056919&group_id=73833&atid=539099]	Fix	NPE at CrawlStateUpdater.java:70 http:/robots.txt	2004-10-29	karl-ia	stack-sf
998275 [http://sourceforge.net/tracker/index.php?func=detail&aid=998275&group_id=73833&atid=539099]	Fix	doc security considerations	2004-07-26	gojomo	gojomo
1549587 [http://sourceforge.net/tracker/index.php?func=detail&aid=1549587&group_id=73833&atid=539099]	Fix	[jdk1.6] ComplexType#toString infinite loop	2006-08-30	stack-sf	stack-sf
1543751 [http://sourceforge.net/tracker/index.php?func=detail&aid=1543751&group_id=73833&atid=539099]	Fix	ConcurrentModificationException in web UI frontier report	2006-08-20	karl-ia	gojomo
1522108 [http://sourceforge.net/tracker/index.php?func=detail&aid=1522108&group_id=73833&atid=539099]	Fix	LinksScoper scope-embedded-links inconsistent/confusing	2006-07-13	gojomo	gojomo
1521563 [http://sourceforge.net/tracker/index.php?func=detail&aid=1521563&group_id=73833&atid=539099]	Fix	UURIFactory '//' collapsing overeager	2006-07-12	karl-ia	gojomo
1519055 [http://sourceforge.net/tracker/index.php?func=detail&aid=1519055&group_id=73833&atid=53	Fix	queued count wrong with retired queues; crawl doesn't end	2006-07-07	karl-ia	gojomo

ID	Type	Summary	Open Date	By	Filer
9099]					
1469517 [http://sourceforge.net/tracker/index.php?func=detail&aid=1469517&group_id=73833&atid=539099]	Fix	ARCWriterPool not fair to threads	2006-04-12	gojomo	gojomo
1379040 [http://sourceforge.net/tracker/index.php?func=detail&aid=1379040&group_id=73833&atid=539099]	Fix	regex for mid-fetch filter not being stored in crawl order	2005-12-12	gojomo	nobody
1550797 [http://sourceforge.net/tracker/index.php?func=detail&aid=1550797&group_id=73833&atid=539099]	Fix	Stripping sessionid can leave behind doubled ampersands	2006-09-01	stack-sf	stack-sf
1541645 [http://sourceforge.net/tracker/index.php?func=detail&aid=1541645&group_id=73833&atid=539099]	Fix	excessive WakeTask may be scheduled	2006-08-16	gojomo	gojomo
1534925 [http://sourceforge.net/tracker/index.php?func=detail&aid=1534925&group_id=73833&atid=539099]	Fix	Remove MirrorJNDI. Its GPL	2006-08-04	stack-sf	stack-sf
1517693 [http://sourceforge.net/tracker/index.php?func=detail&aid=1517693&group_id=73833&atid=539099]	Fix	[extractorhtml] Passes through entity-encodings	2006-07-05	stack-sf	stack-sf
1516354 [http://sourceforge.net/tracker/index.php?func=detail&aid=1516	Fix	Job's crawl report link produces report for different job	2006-07-03	nobody	fmccown

ID	Type	Summary	Open Date	By	Filer
354&group_id=73833&atid=539099]					
1511609 [http://sourceforge.net/tracker/index.php?func=detail&aid=1511609&group_id=73833&atid=539099]	Fix	Browsers tolerate newlines in URLs, Heritrix doesn't	2006-06-23	nobody	stack-sf
1507554 [http://sourceforge.net/tracker/index.php?func=detail&aid=1507554&group_id=73833&atid=539099]	Fix	Values from dropdown getting tacked on for next hit.	2006-06-16	stack-sf	nobody
1503781 [http://sourceforge.net/tracker/index.php?func=detail&aid=1503781&group_id=73833&atid=539099]	Fix	[jmx] Add rebind to JNDI	2006-06-09	nobody	stack-sf
1490806 [http://sourceforge.net/tracker/index.php?func=detail&aid=1490806&group_id=73833&atid=539099]	Fix	hangs with queued documents not being assigned to queues	2006-05-18	nobody	pandae
1489155 [http://sourceforge.net/tracker/index.php?func=detail&aid=1489155&group_id=73833&atid=539099]	Fix	httpclient list of proto-factories is static	2006-05-15	stack-sf	stack-sf
1479727 [http://sourceforge.net/tracker/index.php?func=detail&aid=1479727&group_id=73833&atid=539099]	Fix	Non-serializable class ARCReader contains Exception	2006-05-01	stack-sf	lars_clausen
1469739 [http://sourceforge.net/tracker/in	Fix	escapeJavaScript should escape HTML	2006-04-13	paul_jack	pandae

ID	Type	Summary	Open Date	By	Filer
dex.php?func=detail&aid=1469739&group_id=73833&atid=539099]		problem characters			

11. Release 1.8.0 - 2006-05-05

Release 1.8.0 adds a number of minor improvements and fixes. Most notably, checkpointing can now be achieved with a single command (with the requisite pause/resume done automatically), and all URIs fetched may be tagged with the original seed URI from which they were discovered. (This source URI information is both in the crawl.log and a new 'source-report.txt' report available among the disk file reports.)

We expect release 1.8.0 to be the last release officially supported on JDK 1.4.x ("Java 2") Java; future releases will require JDK 1.5.x ("Java 5") Java facilities.

11.1. Known Limitations/Issues

11.1.1. java.io.IOException: No locks available

BDB-JE will complain 'No locks available' when crawler is being built/run on an NFS mount. Workaround is to locate the 'state' directory on a non-NFS-mounted volume.

11.1.2. "Channel closed, may be due to thread interrupt"

An error with this message has been observed intermittently when running on the Sun Java 6 ("mustang") beta JVM ("-beta2-b81"). A forthcoming fix from Sleepycat for BDB-JE may be necessary to resolve this issue.

11.2. Changes

11.2.1. Progress Statistics Log

The format of progress statistics' state-change log messages have been modified. State-change messages now have a tail that adds some context explaining why we're pausing, etc. Note, we will be adding originator of the status-change event to the progress statistics log post 1.8.0 -- i.e. whether event came of JMX or via the UI -- so be prepared for more progress log changes.

11.2.2. Checkpoints

Now when you ask to checkpoint a running crawl, it will manage for you the pause, checkpoint, and resume cycle (If paused when checkpoint is invoked, the crawler will be set back into a paused state upon checkpoint completion).

Checkpoints made with 1.6.0 software cannot be recovered with 1.8.0 software. Core classes such as CrawlController have changed so their serialized representation as part of a checkpoint has also changed (We have not done the work to deserialize earlier versions of core classes serialized as part of a checkpoint).

Table 4. All Tracked Changes

ID	Type	Summary	Open Date	By	Filer
1440656 [http://sourceforge.net/tracker/index.php?func=detail&aid=1440656&group_id=73833&atid=539099]	Fix	upping total budget doesn't update/unretire queues	2006-02-28	karl-ia	gojomo
1482761 [http://sourceforge.net/tracker/index.php?func=detail&aid=1482761&group_id=73833&atid=539099]	Fix	BDB Adler32 gc-lock OOME risk	2006-05-05	stack-sf	gojomo
1371195 [http://sourceforge.net/tracker/index.php?func=detail&aid=1371195&group_id=73833&atid=539099]	Fix	[jmx] Make downloaded data count have constant units	2005-12-01	stack-sf	stack-sf
1371326 [http://sourceforge.net/tracker/index.php?func=detail&aid=1371326&group_id=73833&atid=539099]	Fix	refactor/compact QuotaEnforcer code	2005-12-01	stack-sf	gojomo
1379208 [http://sourceforge.net/tracker/index.php?func=detail&aid=1379208&group_id=73833&atid=539099]	Fix	crawl report/hosts-report stats leave out robots.txt	2005-12-12	gojomo	gojomo
1415940 [http://sourceforge.net/tracker/index.php?func=detail&aid=1415940&group_id=73833&atid=539099]	Fix	Failed deregistration of container with jndi	2006-01-26	stack-sf	stack-sf
1415942 [http://sourceforge.net/tracker/in	Fix	When multiple instances, there's always a	2006-01-26	stack-sf	stack-sf

ID	Type	Summary	Open Date	By	Filer
dex.php?func=detail&aid=1415942&group_id=73833&atid=539099]		runt in the litter			
1417062 [http://sourceforge.net/tracker/index.php?func=detail&aid=1417062&group_id=73833&atid=539099]	Fix	JMX get alert by index broken.	2006-01-27	stack-sf	stack-sf
1419272 [http://sourceforge.net/tracker/index.php?func=detail&aid=1419272&group_id=73833&atid=539099]	Fix	Corrupt job.state files obstruct crawl resumption	2006-01-30	stack-sf	stack-sf
1442207 [http://sourceforge.net/tracker/index.php?func=detail&aid=1442207&group_id=73833&atid=539099]	Fix	stop alerts 'line in seed file ignored' for mixed seed/surt	2006-03-02	gojomo	ia_igor
1462407 [http://sourceforge.net/tracker/index.php?func=detail&aid=1462407&group_id=73833&atid=539099]	Fix	IllegalArgumentException adding to source host report	2006-03-31	stack-sf	stack-sf
1465369 [http://sourceforge.net/tracker/index.php?func=detail&aid=1465369&group_id=73833&atid=539099]	Fix	make_reports.pl outdated, broken	2006-04-05	gojomo	gojomo
1475730 [http://sourceforge.net/tracker/index.php?func=detail&aid=1475730&group_id=73833&atid=539099]	Fix	OnHostsDecideRule/ OnDomainsDecideRule not adding seed SURTs	2006-04-24	gojomo	gojomo
1475638	Fix	Robots.txt ig-	2006-04-24	gojomo	stack-sf

ID	Type	Summary	Open Date	By	Filer
[http://sourceforge.net/tracker/index.php?func=detail&aid=1475638&group_id=73833&atid=539099]		nored if 206/203 Status Code			
1395637 [http://sourceforge.net/tracker/index.php?func=detail&aid=1395637&group_id=73833&atid=539099]	Fix	crawl.log entires do not reflect 'no space left' error	2006-01-02	karl-ia	ia_igor
1400646 [http://sourceforge.net/tracker/index.php?func=detail&aid=1400646&group_id=73833&atid=539099]	Fix	ExtractorHTML/ExtractorJS 'hang' on many-backslash input	2006-01-09	karl-ia	gojomo
1404316 [http://sourceforge.net/tracker/index.php?func=detail&aid=1404316&group_id=73833&atid=539099]	Fix	ExtractorCSS does not resolve relative URIs against BASE	2006-01-12	karl-ia	ia_igor
1392104 [http://sourceforge.net/tracker/index.php?func=detail&aid=1392104&group_id=73833&atid=539099]	Fix	ExtractorJS NPE doing speculative extraction	2005-12-28	karl-ia	stack-sf
1387423 [http://sourceforge.net/tracker/index.php?func=detail&aid=1387423&group_id=73833&atid=539102]	Add	[arcreader] Fetch records and iterate remote ARCs	2005-12-21	stack-sf	stack-sf
1371178 [http://sourceforge.net/tracker/index.php?func=detail&aid=1371178&group_id=73833&atid=539102]	Add	[jmx] Add name of heritrix 'host' as att	2005-12-01	stack-sf	stack-sf

ID	Type	Summary	Open Date	By	Filer
1233079 [http://sourceforge.net/tracker/index.php?func=detail&aid=1233079&group_id=73833&atid=539102]	Add	replace util.concurrent with java.util.concurrent	2005-07-05	gojomo	gojomo
1371202 [http://sourceforge.net/tracker/index.php?func=detail&aid=1371202&group_id=73833&atid=539102]	Add	[jmx] Regularize crawl end state messages	2005-12-01	stack-sf	stack-sf
1365804 [http://sourceforge.net/tracker/index.php?func=detail&aid=1365804&group_id=73833&atid=539102]	Add	JmxUtils.getOpenType() must handle Doubles	2005-11-24	stack-sf	nobody
1374947 [http://sourceforge.net/tracker/index.php?func=detail&aid=1374947&group_id=73833&atid=539102]	Add	[jmx] progress statistics as notification	2005-12-06	stack-sf	stack-sf
1388275 [http://sourceforge.net/tracker/index.php?func=detail&aid=1388275&group_id=73833&atid=539102]	Add	[contrib] Preselector ATTR_ALLOW_BY_REGEXP	2005-12-22	stack-sf	stack-sf
1393254 [http://sourceforge.net/tracker/index.php?func=detail&aid=1393254&group_id=73833&atid=539102]	Add	'total' bytes/fetches quota options in QuotaEnforcer	2005-12-29	gojomo	gojomo
1119608 [http://sourceforge.net/tracker/index.php?func=detail&aid=1119608&group_id=73833&atid=53	Add	Carry forward (& log) 'originating URL/seed' for all URLs	2005-02-09	gojomo	gojomo

ID	Type	Summary	Open Date	By	Filer
9102]					
1358617 [http://sourceforge.net/tracker/index.php?func=detail&aid=1358617&group_id=73833&atid=539102]	Add	Add destroy to JMX API	2005-11-16	karl-ia	stack-sf
1445970 [http://sourceforge.net/tracker/index.php?func=detail&aid=1445970&group_id=73833&atid=539102]	Add	New "seed source report" of # of URLs per host per source	2006-03-08	karl-ia	stack-sf
1436290 [http://sourceforge.net/tracker/index.php?func=detail&aid=1436290&group_id=73833&atid=539102]	Add	improve surt docs; esp 'surts-source-file' syntax	2006-02-21	nobody	gojomo
1302207 [http://sourceforge.net/tracker/index.php?func=detail&aid=1302207&group_id=73833&atid=539102]	Add	unattended checkpointing	2005-09-23	karl-ia	gojomo

12. Release 1.6.0 - 2005-12-01

Release 1.6.0 offers improved remote control and monitoring via JMX, a crawl-checkpointing facility, and experimental support for bloom filter already-included testing, partitioning a crawl across multiple independent crawlers, and per-host/domain/queue-grouping collection quotas. Performance and stability in large crawls is also improved. Among tracked issues, it includes 39 requested enhancements and fixes 96 reported bugs.

12.1. Known Limitations/Issues

12.1.1. java.io.IOException: No locks available

BDB will complain 'No locks available' when crawler is being built/run on an NFS mount. Workaround is not run on an NFS-mounted volume.

12.1.2. OutOfMemoryError in 64bit JVMs

BDB 2.0.90 can overgrow its intended cache size due to a misestimation of instance sizes under 64bit

Java VMs, which may be a major contributor to early Heritrix OutOfMemoryError problems on 64bit systems. A workaround is to cut the assigned percentage by 1/3 to 1/2. For example, change the 'bdb-cache-percent' setting to '40' or '30' (instead of the default 60% when no value is set here).

12.2. Changes

12.2.1. Postselector

The Postselector has been refactored out of existence. Its responsibilities have been parcelled out to two new Processors: LinksScoper and FrontierScheduler. LinksScoper is responsible for scope checking of extracted links. FrontierScheduler does the scheduling of URIs with the Frontier.

This change was done to allow introduction of processors between scope checking and Frontier scheduling steps.

Because of this change, order files from 1.4.0 Heritrix or before will need to be updated -- Postselector references replaced by LinkScoper and FrontierScheduler references -- before they can be used with Heritrix 1.6.0 (Referencing a non-existent Postselector in an order file usually shows as -50 fetch status in crawl.log).

12.2.2. Web Console

The layout and terminology of the web Console and header have been changed, and new readouts added. Most notably, "Crawler Status" and "Job Status" information have been moved to separate boxes, with the controls for each at the top of their respective boxes, near the current status information. Also, the "Crawling"/"Stopped" distinction in the crawler -- whether available pending jobs would be started as possible -- has been renamed "Crawling Jobs"/"Holding Jobs" for clarity.

Table 5. All Tracked Changes

ID	Type	Summary	Open Date	By	Filer
806831 [http://sourceforge.net/tracker/index.php?func=detail&aid=806831&group_id=73833&atid=539102]	Add	XMLExtractor (XML/RSS)	2003-09-15	gojomo	gojomo
983051 [http://sourceforge.net/tracker/index.php?func=detail&aid=983051&group_id=73833&atid=539102]	Add	annotate what robots.txt would have precluded	2004-06-30	karl-ia	gojomo
1069331 [http://sourceforge.net/tracker/index.php?func=detail&aid=1069331&group_id=	Add	hold paused crawl at 'end', allowing all in-progress ops	2004-11-19	karl-ia	gojomo

ID	Type	Summary	Open Date	By	Filer
73833&atid=539102]					
1081774 [http://sourceforge.net/tracker/index.php?func=detail&aid=1081774&group_id=73833&atid=539102]	Add	need way to delete overrides	2004-12-08	karl-ia	gojomo
1104696 [http://sourceforge.net/tracker/index.php?func=detail&aid=1104696&group_id=73833&atid=539102]	Add	Confusion: CrawlController and CrawlJob States	2005-01-18	nobody	stack-sf
1108006 [http://sourceforge.net/tracker/index.php?func=detail&aid=1108006&group_id=73833&atid=539102]	Add	alerts should show current processor	2005-01-23	gojomo	gojomo
1108520 [http://sourceforge.net/tracker/index.php?func=detail&aid=1108520&group_id=73833&atid=539102]	Add	SURT needs facelift	2005-01-24	gojomo	stack-sf
1119616 [http://sourceforge.net/tracker/index.php?func=detail&aid=1119616&group_id=73833&atid=539102]	Add	Decompose Postselector to Scoping and Scheduling components	2005-02-09	stack-sf	gojomo
1122692 [http://sourceforge.net/tracker/index.php?func=detail&aid=1122692&group_id=73833&atid=539102]	Add	[contribution] New fixed number of queues policy	2005-02-14	stack-sf	stack-sf
1173597 [http://sourceforge.net/tracker/index.php?func=d	Add	jmx api additions	2005-03-30	stack-sf	stack-sf

ID	Type	Summary	Open Date	By	Filer
etail&aid=1173597&group_id=73833&atid=539102]					
1176934 [http://sourceforge.net/tracker/index.php?func=detail&aid=1176934&group_id=73833&atid=539102]	Add	[contrib] Generalize/Refactor BDB Frontier	2005-04-05	stack-sf	ck-heritrix
1180630 [http://sourceforge.net/tracker/index.php?func=detail&aid=1180630&group_id=73833&atid=539102]	Add	[contrib] UI stacktrace dump (Depends on JDK150)	2005-04-11	stack-sf	ck-heritrix
1183376 [http://sourceforge.net/tracker/index.php?func=detail&aid=1183376&group_id=73833&atid=539102]	Add	Post 1.4 Deprecate filter scope and remove post 1.6.	2005-04-14	stack-sf	stack-sf
1190974 [http://sourceforge.net/tracker/index.php?func=detail&aid=1190974&group_id=73833&atid=539102]	Add	Quick resume without real recovery / Checkpointing	2005-04-27	karl-ia	ck-heritrix
1196602 [http://sourceforge.net/tracker/index.php?func=detail&aid=1196602&group_id=73833&atid=539102]	Add	[contrib] Show estimated remaining time	2005-05-06	stack-sf	ck-heritrix
1200205 [http://sourceforge.net/tracker/index.php?func=detail&aid=1200205&group_id=73833&atid=539102]	Add	add 'exhausted' queue count to frontier report	2005-05-11	gojomo	gojomo
1204644 [http://sourcefor	Add	add 'memory used' to pro-	2005-05-18	kristinn_sig	gojomo

ID	Type	Summary	Open Date	By	Filer
ge.net/tracker/index.php?func=detail&aid=1204644&group_id=73833&atid=539102]		gress-statistics.log			
1205583 [http://sourceforge.net/tracker/index.php?func=detail&aid=1205583&group_id=73833&atid=539102]	Add	add CandidateURI parameter to UriUniqFilter.forget()	2005-05-20	stack-sf	ck-heritrix
1207866 [http://sourceforge.net/tracker/index.php?func=detail&aid=1207866&group_id=73833&atid=539102]	Add	[contrib] ThreadLocal-version of TextUtil.getMatcher	2005-05-24	gojomo	ck-heritrix
1207898 [http://sourceforge.net/tracker/index.php?func=detail&aid=1207898&group_id=73833&atid=539102]	Add	[contrib] WorkQueueFrontier: Store allQueues in RAM if poss.	2005-05-24	stack-sf	ck-heritrix
1208293 [http://sourceforge.net/tracker/index.php?func=detail&aid=1208293&group_id=73833&atid=539102]	Add	List based URIRegExprFilter	2005-05-25	kristinn_sig	kristinn_sig
1208510 [http://sourceforge.net/tracker/index.php?func=detail&aid=1208510&group_id=73833&atid=539102]	Add	[rfe-contrib] Add Stacktrace dump to ToeThread.report()	2005-05-25	stack-sf	ck-heritrix
1208747 [http://sourceforge.net/tracker/index.php?func=detail&aid=1208747&group_id=73833&atid=539102]	Add	CrawlURI serialization bloated; should be slimmed	2005-05-25	gojomo	gojomo

ID	Type	Summary	Open Date	By	Filer
1208757 [http://sourceforge.net/tracker/index.php?func=detail&aid=1208757&group_id=73833&atid=539102]	Add	Cookies are thread traffic jam and memory hog	2005-05-25	gojomo	stack-sf
1208770 [http://sourceforge.net/tracker/index.php?func=detail&aid=1208770&group_id=73833&atid=539102]	Add	garbage hot spot: Serial-Binding & FastOutput-Stream.bump()	2005-05-25	gojomo	gojomo
1211217 [http://sourceforge.net/tracker/index.php?func=detail&aid=1211217&group_id=73833&atid=539102]	Add	[contrib] Add debugging aid for BDB RuntimeExcep-tionWrapper	2005-05-30	stack-sf	ck-heritrix
1217854 [http://sourceforge.net/tracker/index.php?func=detail&aid=1217854&group_id=73833&atid=539102]	Add	seed report of redirect should show where to	2005-06-09	karl-ia	gojomo
1222764 [http://sourceforge.net/tracker/index.php?func=detail&aid=1222764&group_id=73833&atid=539102]	Add	Rotation of crawl logs	2005-06-17	karl-ia	stack-sf
1223840 [http://sourceforge.net/tracker/index.php?func=detail&aid=1223840&group_id=73833&atid=539102]	Add	BdbWorkQueue origins should be based on full classKey	2005-06-19	gojomo	gojomo
1225597 [http://sourceforge.net/tracker/index.php?func=detail&aid=1225597&group_id=73833&atid=53	Add	Expose the bdb je 2.0 jmx inter-face	2005-06-22	karl-ia	stack-sf

ID	Type	Summary	Open Date	By	Filer
9102]					
1225729 [http://sourceforge.net/tracker/index.php?func=detail&aid=1225729&group_id=73833&atid=539102]	Add	new alreadyIncluded option: Bloom filter based	2005-06-22	gojomo	gojomo
1254560 [http://sourceforge.net/tracker/index.php?func=detail&aid=1254560&group_id=73833&atid=539102]	Add	Add to queue-assignment-policy without compile	2005-08-08	karl-ia	stack-sf
1260360 [http://sourceforge.net/tracker/index.php?func=detail&aid=1260360&group_id=73833&atid=539102]	Add	More than one Heritrix instance in a JVM instance	2005-08-15	karl-ia	stack-sf
1261506 [http://sourceforge.net/tracker/index.php?func=detail&aid=1261506&group_id=73833&atid=539102]	Add	Multimachine Crawl Splitter Processor	2005-08-16	gojomo	stack-sf
1262665 [http://sourceforge.net/tracker/index.php?func=detail&aid=1262665&group_id=73833&atid=539102]	Add	add dummy items at heads of BDB queues for performance	2005-08-17	gojomo	gojomo
1302182 [http://sourceforge.net/tracker/index.php?func=detail&aid=1302182&group_id=73833&atid=539102]	Add	IP geolocation based scoping	2005-09-23	karl-ia	gojomo
1302208 [http://sourceforge.net/tracker/index.php?func=detail&aid=1302	Add	per-crawler 'load' summary numbers	2005-09-23	karl-ia	gojomo

ID	Type	Summary	Open Date	By	Filer
208&group_id=73833&atid=539102]					
1325123 [http://sourceforge.net/tracker/index.php?func=detail&aid=1325123&group_id=73833&atid=539102]	Add	Checkpointing fixes/ improvements	2005-10-12	stack-sf	stack-sf
1329725 [http://sourceforge.net/tracker/index.php?func=detail&aid=1329725&group_id=73833&atid=539102]	Add	[uuri] When 'generous' mode, don't encode curly-brackets	2005-10-18	stack-sf	stack-sf
965622 [http://sourceforge.net/tracker/index.php?func=detail&aid=965622&group_id=73833&atid=539099]	Fix	Serious error during crawling did not produce an alert	2004-06-03	gojomo	kristinn_sig
1000338 [http://sourceforge.net/tracker/index.php?func=detail&aid=1000338&group_id=73833&atid=539099]	Fix	[UURI] escaped absolute path not valid	2004-07-29	nobody	stack-sf
1002356 [http://sourceforge.net/tracker/index.php?func=detail&aid=1002356&group_id=73833&atid=539099]	Fix	timing issue on crawl-start & "run time" stat	2004-08-02	gojomo	gojomo
1059126 [http://sourceforge.net/tracker/index.php?func=detail&aid=1059126&group_id=73833&atid=539099]	Fix	Other than default seeds path is not used	2004-11-02	nobody	stack-sf
1060517 [http://sourceforge.net/tracker/in	Fix	hard-coding of job dir in state.job makes	2004-11-04	nobody	stack-sf

ID	Type	Summary	Open Date	By	Filer
dex.php?func=detail&aid=1060517&group_id=73833&atid=539099]		moves awkward			
1062727 [http://sourceforge.net/tracker/index.php?func=detail&aid=1062727&group_id=73833&atid=539099]	Fix	angle-brackets in URIs thwart frontier report	2004-11-08	kristinn_sig	gojomo
1065413 [http://sourceforge.net/tracker/index.php?func=detail&aid=1065413&group_id=73833&atid=539099]	Fix	[uuri] '$' in path gets scheduled, spawns queueing error	2004-11-12	gojomo	stack-sf
1080926 [http://sourceforge.net/tracker/index.php?func=detail&aid=1080926&group_id=73833&atid=539099]	Fix	reducing max-toe-threads has no effect	2004-12-07	gojomo	gojomo
1083427 [http://sourceforge.net/tracker/index.php?func=detail&aid=1083427&group_id=73833&atid=539099]	Fix	Text incorrect in WUI when create new profile	2004-12-11	nobody	zhousp
1090564 [http://sourceforge.net/tracker/index.php?func=detail&aid=1090564&group_id=73833&atid=539099]	Fix	max-trans-hops=0 generates -63 in crawl.log	2004-12-23	gojomo	stack-sf
1090916 [http://sourceforge.net/tracker/index.php?func=detail&aid=1090916&group_id=73833&atid=539099]	Fix	RIS still open; ThreadLocal-Connection-Manager async close	2004-12-24	karl-ia	stack-sf
1113410	Fix	NPE	2005-01-31	gojomo	stack-sf

ID	Type	Summary	Open Date	By	Filer
[http://sourceforge.net/tracker/index.php?func=detail&aid=1113410&group_id=73833&atid=539099]		readalert_jsp._jspService			
1116456 [http://sourceforge.net/tracker/index.php?func=detail&aid=1116456&group_id=73833&atid=539099]	Fix	ARCWriter length is wrong (But coherent gzip record)	2005-02-04	karl-ia	stack-sf
1119644 [http://sourceforge.net/tracker/index.php?func=detail&aid=1119644&group_id=73833&atid=539099]	Fix	frontier report Concurrent-Modifica-tionException	2005-02-09	gojomo	frodobay
1122836 [http://sourceforge.net/tracker/index.php?func=detail&aid=1122836&group_id=73833&atid=539099]	Fix	Localize Stack-OverflowError in Extractors	2005-02-14	nobody	gojomo
1181892 [http://sourceforge.net/tracker/index.php?func=detail&aid=1181892&group_id=73833&atid=539099]	Fix	Aggressive ex-traction of `for' attributes	2005-04-12	karl-ia	ia_igor
1187973 [http://sourceforge.net/tracker/index.php?func=detail&aid=1187973&group_id=73833&atid=539099]	Fix	NPE in FetchDNS, caused by UURI	2005-04-22	nobody	ck-heritrix
1192029 [http://sourceforge.net/tracker/index.php?func=detail&aid=1192029&group_id=73833&atid=539099]	Fix	OOME guard against pages of thousands of links	2005-04-28	gojomo	stack-sf

ID	Type	Summary	Open Date	By	Filer
1195312 [http://sourceforge.net/tracker/index.php?func=detail&aid=1195312&group_id=73833&atid=539099]	Fix	WaitEvaluators accidentally removed from Processors.options	2005-05-04	nobody	stack-sf
1196594 [http://sourceforge.net/tracker/index.php?func=detail&aid=1196594&group_id=73833&atid=539099]	Fix	minor CSS typo, wrong text font.	2005-05-06	nobody	ck-heritrix
1196630 [http://sourceforge.net/tracker/index.php?func=detail&aid=1196630&group_id=73833&atid=539099]	Fix	Build w/ 150 jdk won't run under 14x.	2005-05-06	stack-sf	stack-sf
1200957 [http://sourceforge.net/tracker/index.php?func=detail&aid=1200957&group_id=73833&atid=539099]	Fix	Web UI recovery mangles paths	2005-05-12	karl-ia	karl-ia
1203235 [http://sourceforge.net/tracker/index.php?func=detail&aid=1203235&group_id=73833&atid=539099]	Fix	can't change cost policy mid-crawl	2005-05-16	gojomo	gojomo
1203588 [http://sourceforge.net/tracker/index.php?func=detail&aid=1203588&group_id=73833&atid=539099]	Fix	BdbFrontier, serious exception - LatchNotHeldException	2005-05-17	karl-ia	kristinn_sig
1203958 [http://sourceforge.net/tracker/index.php?func=detail&aid=1203958&group_id=73833&atid=53	Fix	JMX remote command 'stop' leaves zombie crawler	2005-05-17	stack-sf	karl-ia

ID	Type	Summary	Open Date	By	Filer
9099]					
1204643 [http://sourceforge.net/tracker/index.php?func=detail&aid=1204643&group_id=73833&atid=539099]	Fix	uri-errors.log: old timestamps, too many errors	2005-05-18	gojomo	gojomo
1204667 [http://sourceforge.net/tracker/index.php?func=detail&aid=1204667&group_id=73833&atid=539099]	Fix	'custom' robots policy doesn't work	2005-05-18	gojomo	gojomo
1204931 [http://sourceforge.net/tracker/index.php?func=detail&aid=1204931&group_id=73833&atid=539099]	Fix	NPE when viewing crawl report	2005-05-19	karl-ia	kristinn_sig
1207320 [http://sourceforge.net/tracker/index.php?func=detail&aid=1207320&group_id=73833&atid=539099]	Fix	[arcwriter] Record w/ empty body on OOME	2005-05-23	stack-sf	stack-sf
1207378 [http://sourceforge.net/tracker/index.php?func=detail&aid=1207378&group_id=73833&atid=539099]	Fix	seeds listed without scheme, but with path, being ignored	2005-05-23	karl-ia	ia_igor
1208804 [http://sourceforge.net/tracker/index.php?func=detail&aid=1208804&group_id=73833&atid=539099]	Fix	CachedBdbMap NPE killing off threads.	2005-05-25	karl-ia	stack-sf
1209046 [http://sourceforge.net/tracker/index.php?func=detail&aid=1209	Fix	Failed URIs should be 'free' (no cost against queue budget)	2005-05-26	gojomo	kristinn_sig

ID	Type	Summary	Open Date	By	Filer
046&group_id=73833&atid=539099]					
1209665 [http://sourceforge.net/tracker/index.php?func=detail&aid=1209665&group_id=73833&atid=539099]	Fix	ReplayCharSequenceFactory: Unexpected response body offset	2005-05-27	karl-ia	ck-heritrix
1212377 [http://sourceforge.net/tracker/index.php?func=detail&aid=1212377&group_id=73833&atid=539099]	Fix	URIException in deserialization, post CrawlURI slimming	2005-05-31	gojomo	gojomo
1213095 [http://sourceforge.net/tracker/index.php?func=detail&aid=1213095&group_id=73833&atid=539099]	Fix	UURI handling of inconsistent escaping makes broken instance	2005-06-01	karl-ia	gojomo
1214478 [http://sourceforge.net/tracker/index.php?func=detail&aid=1214478&group_id=73833&atid=539099]	Fix	ThreadLocalHttpConnectionManager starts a non-daemon Thread	2005-06-03	nobody	ck-heritrix
1216633 [http://sourceforge.net/tracker/index.php?func=detail&aid=1216633&group_id=73833&atid=539099]	Fix	recovery fills heritrix_out with "Relative URI but no base..	2005-06-07	gojomo	gojomo
1217290 [http://sourceforge.net/tracker/index.php?func=detail&aid=1217290&group_id=73833&atid=539099]	Fix	Non-canonical seed URLs need better reporting	2005-06-08	karl-ia	karl-ia
1218019 [http://sourceforge.net/tracker/in	Fix	GzippedInputStream class is not thread-safe	2005-06-10	nobody	ck-heritrix

ID	Type	Summary	Open Date	By	Filer
dex.php?func=detail&aid=1218019&group_id=73833&atid=539099]					
1218037 [http://sourceforge.net/tracker/index.php?func=detail&aid=1218037&group_id=73833&atid=539099]	Fix	CookieSpec interface modification breaks IgnoreCookiesSpec	2005-06-10	gojomo	ck-heritrix
1218283 [http://sourceforge.net/tracker/index.php?func=detail&aid=1218283&group_id=73833&atid=539099]	Fix	ARCReader: Bad URI escaping for tab character	2005-06-10	nobody	ck-heritrix
1218958 [http://sourceforge.net/tracker/index.php?func=detail&aid=1218958&group_id=73833&atid=539099]	Fix	odd preencoded URIs generate error on deserialization	2005-06-11	nobody	gojomo
1219259 [http://sourceforge.net/tracker/index.php?func=detail&aid=1219259&group_id=73833&atid=539099]	Fix	broad crawls slow; most threads stuck retrying missing sites	2005-06-12	gojomo	gojomo
1219262 [http://sourceforge.net/tracker/index.php?func=detail&aid=1219262&group_id=73833&atid=539099]	Fix	'treat seed redirects as new seeds' not working	2005-06-12	karl-ia	gojomo
1219486 [http://sourceforge.net/tracker/index.php?func=detail&aid=1219486&group_id=73833&atid=539099]	Fix	no rule for decidingscope to always crawl seeds	2005-06-12	gojomo	gojomo
1219715	Fix	[patch] Signa-	2005-06-13	nobody	ck-heritrix

ID	Type	Summary	Open Date	By	Filer
[http://sourceforge.net/tracker/index.php?func=detail&aid=1219715&group_id=73833&atid=539099]		ture change broke BucketQueueAssignmentPolicy			
1219854 [http://sourceforge.net/tracker/index.php?func=detail&aid=1219854&group_id=73833&atid=539099]	Fix	NPE je-2.0 entryToObject(SerialBinding.java:82)	2005-06-13	karl-ia	stack-sf
1220714 [http://sourceforge.net/tracker/index.php?func=detail&aid=1220714&group_id=73833&atid=539099]	Fix	ExtractorHTML excessive temp strings / OOM	2005-06-14	karl-ia	gojomo
1221570 [http://sourceforge.net/tracker/index.php?func=detail&aid=1221570&group_id=73833&atid=539099]	Fix	reports (web ui and to disk) don't scale	2005-06-15	gojomo	gojomo
1222229 [http://sourceforge.net/tracker/index.php?func=detail&aid=1222229&group_id=73833&atid=539099]	Fix	unicode/idn domain names fail (seeds and more?)- punycode	2005-06-16	karl-ia	gojomo
1222360 [http://sourceforge.net/tracker/index.php?func=detail&aid=1222360&group_id=73833&atid=539099]	Fix	strip-www canon. rule causes failed crawl of netarkivet.dk	2005-06-16	karl-ia	stack-sf
1224531 [http://sourceforge.net/tracker/index.php?func=detail&aid=1224531&group_id=73833&atid=539099]	Fix	'@' in URI path confuses SURT (bad queues, scope probs, etc)	2005-06-20	karl-ia	gojomo

ID	Type	Summary	Open Date	By	Filer
1226365 [http://sourceforge.net/tracker/index.php?func=detail&aid=1226365&group_id=73833&atid=539099]	Fix	DomainSensitiveFrontier broken by uri-included-structure	2005-06-23	karl-ia	stack-sf
1226387 [http://sourceforge.net/tracker/index.php?func=detail&aid=1226387&group_id=73833&atid=539099]	Fix	BdbUriUniqFilter URI fingerprint somewhat collision prone	2005-06-23	karl-ia	gojomo
1226707 [http://sourceforge.net/tracker/index.php?func=detail&aid=1226707&group_id=73833&atid=539099]	Fix	CandidateURI serialization 'decodes' UURI	2005-06-23	karl-ia	stack-sf
1230180 [http://sourceforge.net/tracker/index.php?func=detail&aid=1230180&group_id=73833&atid=539099]	Fix	nonstandard port URIs fail with '-50': Surt queue policy bug	2005-06-30	karl-ia	gojomo
1230188 [http://sourceforge.net/tracker/index.php?func=detail&aid=1230188&group_id=73833&atid=539099]	Fix	DNS prereq problems (-50 fails/repeats): calculateInsertKey	2005-06-30	karl-ia	gojomo
1231123 [http://sourceforge.net/tracker/index.php?func=detail&aid=1231123&group_id=73833&atid=539099]	Fix	IdentityCachingMapTest FAILED: cache not cleared appropriate	2005-07-01	gojomo	stack-sf
1232402 [http://sourceforge.net/tracker/index.php?func=detail&aid=1232402&group_id=73833&atid=53	Fix	IdentityCachingMapTest fails on fedora	2005-07-04	nobody	stack-sf

ID	Type	Summary	Open Date	By	Filer
9099]					
1232974 [http://sourceforge.net/tracker/index.php?func=detail&aid=1232974&group_id=73833&atid=539099]	Fix	WorkQueueFrontier.kickUpdate ClassCastException / unretiring	2005-07-05	karl-ia	gojomo
1236094 [http://sourceforge.net/tracker/index.php?func=detail&aid=1236094&group_id=73833&atid=539099]	Fix	Possible deadlock situation with ARFrontier	2005-07-11	kristinn_sig	kristinn_sig
1236334 [http://sourceforge.net/tracker/index.php?func=detail&aid=1236334&group_id=73833&atid=539099]	Fix	Cannot set cachePercentage in bdbje JMX bean	2005-07-11	stack-sf	stack-sf
1236635 [http://sourceforge.net/tracker/index.php?func=detail&aid=1236635&group_id=73833&atid=539099]	Fix	Link ClassCastException	2005-07-12	kristinn_sig	kristinn_sig
1239155 [http://sourceforge.net/tracker/index.php?func=detail&aid=1239155&group_id=73833&atid=539099]	Fix	[arcreader] Fails on records that only have headers	2005-07-15	karl-ia	stack-sf
1241851 [http://sourceforge.net/tracker/index.php?func=detail&aid=1241851&group_id=73833&atid=539099]	Fix	Connection reset error with WebSTAR/ 3.0.2 web serve	2005-07-20	karl-ia	ia_igor
1242747 [http://sourceforge.net/tracker/index.php?func=detail&aid=1242	Fix	over-escaping (of '%', etc) compared to browsers	2005-07-21	karl-ia	gojomo

ID	Type	Summary	Open Date	By	Filer
747&group_id=73833&atid=539099]					
1248942 [http://sourceforge.net/tracker/index.php?func=detail&aid=1248942&group_id=73833&atid=539099]	Fix	custom robots.txt NPE in 1.4.0	2005-07-31	gojomo	efc
1249828 [http://sourceforge.net/tracker/index.php?func=detail&aid=1249828&group_id=73833&atid=539099]	Fix	-5000 out-of-scope pre-conditions; -50 failure	2005-08-01	karl-ia	gojomo
1250437 [http://sourceforge.net/tracker/index.php?func=detail&aid=1250437&group_id=73833&atid=539099]	Fix	host with trailing '.' in seeds ruins implied SURT	2005-08-02	gojomo	gojomo
1255137 [http://sourceforge.net/tracker/index.php?func=detail&aid=1255137&group_id=73833&atid=539099]	Fix	Deferred URI should be 'free' (no cost against queue budget)	2005-08-09	karl-ia	gojomo
1257157 [http://sourceforge.net/tracker/index.php?func=detail&aid=1257157&group_id=73833&atid=539099]	Fix	create job: empty seeds box, broken settings tab	2005-08-11	gojomo	gojomo
1266713 [http://sourceforge.net/tracker/index.php?func=detail&aid=1266713&group_id=73833&atid=539099]	Fix	already-seen test not working (1.4.0); canonicalization	2005-08-22	nobody	stack-sf
1276044 [http://sourceforge.net/tracker/in	Fix	Prune classes deprecated in 1.4.0	2005-08-29	nobody	stack-sf

ID	Type	Summary	Open Date	By	Filer
dex.php?func=detail&aid=1276044&group_id=73833&atid=539099]					
1276201 [http://sourceforge.net/tracker/index.php?func=detail&aid=1276201&group_id=73833&atid=539099]	Fix	Notification AFTER reports have been created	2005-08-29	stack-sf	stack-sf
1283492 [http://sourceforge.net/tracker/index.php?func=detail&aid=1283492&group_id=73833&atid=539099]	Fix	mimetype-reports squashes together url-count and byte-count	2005-09-06	gojomo	stack-sf
1284353 [http://sourceforge.net/tracker/index.php?func=detail&aid=1284353&group_id=73833&atid=539099]	Fix	Help links to user and dev manuals 404	2005-09-07	stack-sf	stack-sf
1291274 [http://sourceforge.net/tracker/index.php?func=detail&aid=1291274&group_id=73833&atid=539099]	Fix	StatisticsTracker writes 'not set'. Crawl.log says 'no-type'	2005-09-14	karl-ia	stack-sf
1291305 [http://sourceforge.net/tracker/index.php?func=detail&aid=1291305&group_id=73833&atid=539099]	Fix	mime type report not consistent with crawl.log	2005-09-14	ia_igor	ia_igor
1306421 [http://sourceforge.net/tracker/index.php?func=detail&aid=1306421&group_id=73833&atid=539099]	Fix	ExtractorUniversal skips URI with '_'	2005-09-27	gojomo	gojomo
1323281	Fix	StringIndex-	2005-10-10	gojomo	gojomo

ID	Type	Summary	Open Date	By	Filer
[http://sourceforge.net/tracker/index.php?func=detail&aid=1323281&group_id=73833&atid=539099]		OutOfBounds in SurtAuthorityQueueAssignmentPolicy			
1323287 [http://sourceforge.net/tracker/index.php?func=detail&aid=1323287&group_id=73833&atid=539099]	Fix	BDB ArrayIndexOutOfBoundsException; corrupt queue	2005-10-10	karl-ia	gojomo
1323323 [http://sourceforge.net/tracker/index.php?func=detail&aid=1323323&group_id=73833&atid=539099]	Fix	IDNAException: String too long / fixupDomainLabel in stdout	2005-10-10	karl-ia	gojomo
1323373 [http://sourceforge.net/tracker/index.php?func=detail&aid=1323373&group_id=73833&atid=539099]	Fix	NPE in CrawlJobHandler.startNextJobInternal	2005-10-10	karl-ia	gojomo
1324245 [http://sourceforge.net/tracker/index.php?func=detail&aid=1324245&group_id=73833&atid=539099]	Fix	implied-HTTP seeds with port numbers get -5000 errs	2005-10-11	karl-ia	gojomo
1325304 [http://sourceforge.net/tracker/index.php?func=detail&aid=1325304&group_id=73833&atid=539099]	Fix	'expert' view toggle does inadvertent submit	2005-10-12	karl-ia	gojomo
1332722 [http://sourceforge.net/tracker/index.php?func=detail&aid=1332722&group_id=73833&atid=539099]	Fix	time-based averages wrong after checkpoint	2005-10-19	karl-ia	gojomo

ID	Type	Summary	Open Date	By	Filer
1333669 [http://sourceforge.net/tracker/index.php?func=detail&aid=1333669&group_id=73833&atid=539099]	Fix	delete function in "View or Edit Frontier URIs" Unsupported	2005-10-20	karl-ia	stack-sf
1344265 [http://sourceforge.net/tracker/index.php?func=detail&aid=1344265&group_id=73833&atid=539099]	Fix	Laxuri code encodes tildes.	2005-10-31	karl-ia	stack-sf
1351818 [http://sourceforge.net/tracker/index.php?func=detail&aid=1351818&group_id=73833&atid=539099]	Fix	Investigate spate of recent OOMEs	2005-11-08	gojomo	stack-sf
1357528 [http://sourceforge.net/tracker/index.php?func=detail&aid=1357528&group_id=73833&atid=539099]	Fix	Double JMX registration problem on remote creation	2005-11-15	stack-sf	stack-sf
1358542 [http://sourceforge.net/tracker/index.php?func=detail&aid=1358542&group_id=73833&atid=539099]	Fix	DevUtils.extraInfo shows 2 legend lines, no content	2005-11-16	gojomo	gojomo
1358567 [http://sourceforge.net/tracker/index.php?func=detail&aid=1358567&group_id=73833&atid=539099]	Fix	transient JSP NPE after starting job	2005-11-16	stack-sf	gojomo
1369177 [http://sourceforge.net/tracker/index.php?func=detail&aid=1369177&group_id=73833&atid=53	Fix	NPE in quotaEnforcer, if hostname for URI can't be resolved	2005-11-29	nobody	svc

ID	Type	Summary	Open Date	By	Filer
9099]					
1369619 [http://sourceforge.net/tracker/index.php?func=detail&aid=1369619&group_id=73833&atid=539099]	Fix	NPE in QuotaEnforcer	2005-11-29	gojomo	gojomo
1370743 [http://sourceforge.net/tracker/index.php?func=detail&aid=1370743&group_id=73833&atid=539099]	Fix	properties-based supported/ignored scheme setting broken	2005-12-01	gojomo	gojomo
1370761 [http://sourceforge.net/tracker/index.php?func=detail&aid=1370761&group_id=73833&atid=539099]	Fix	ArithmeticException: / by 0 WorkQueueFrontier.averageDepth	2005-12-01	gojomo	gojomo
1325230 [http://sourceforge.net/tracker/index.php?func=detail-amp;aid=1325230&group_id=73833&atid=539099]	Fix	jmx import of seeds not working	2005-10-12 14:29	karl-ia	stack-sf
1324989 [http://sourceforge.net/tracker/index.php?func=detail-amp;aid=1324989&group_id=73833&atid=539099]	Fix	Queue counts wrong after checkpointing	2005-10-12 09:10	karl-ia	stack-sf
1123230 [http://sourceforge.net/tracker/index.php?func=detail-amp;aid=1123230&group_id=73833&atid=539099]	Fix	Out Of Memory after creating mutliple jobs	2005-02-15 07:51	karl-ia	nobody
1322280 [http://sourcefor	Fix	"Failed getPath" alerts (from Ro-	2005-10-10 02:42	karl-ia	gojomo

ID	Type	Summary	Open Date	By	Filer
ge.net/tracker/index.php?func=detail-amp;aid=1322280&group_id=73833&atid=539099]		botsExclusion-Policy)			
1322264 [http://sourceforge.net/tracker/index.php?func=detail-amp;aid=1322264&group_id=73833&atid=539099]	Fix	NumberFormatException in FetchHTTP.innerProcess	2005-10-10 02:34	karl-ia	gojomo

13. Release 1.4.0 - 2005-04-28

Much improved memory usage, new scoping/filter model, and a new revisiting frontier. Over 90 bugs fixed.

13.1. Known Limitations/Issues

13.1.1. Glibc 2.3.2 and NPTL

NPTL is the 'new' linux threading model. It replaces *linuxthreads* the 'old' model. You can tell you're running NPTL if your java process shows as one process only in the process listing. Wwith linuxthreads, all java threads show as distinct linux processes. Linux threading is integral to glibc.

On rare occasions we've seen the crawler hang without obvious explaination when running with NPTL threading on linux. Doing a thread dump on the hung crawler, one version of the hung crawler has threads waiting to obtain a lock that no one apparently holds. Our reading has these rare, crawl-killing, hangs as a problem in glbc2.3.2 when running with NPTL (NPTL 0.60) (We used to hang frequently but workarounds seem to have mitigated the frequency of lockup making it extremely rare). An upgrade to glibc2.3.3+ seems to do away with these hangs. Glibc2.3.3 has NPTL 0.61. Fedora3 has glibc2.3.4. If an upgrade is not possible -- for example, the new glibc is not currently available for debian -- you can disable NPTL and run with old threads by setting the environment variable `LD_ASSUME_KERNEL=2.4.1` (You can set this environment variable on a per process basis).

NPTL is usually the default threading model on linux and is usually what you want -- threads are more lightweight and java throughput seems to be slightly higher with NPTL enabled. Various are the ways in which you can see which threading model you are using. Do an ldd on the java executable to see what shared libraries its using. Note the location of the glibc shared library. Executing `PATH_TO_GLIBC/lib.so.6`, usually `/lib/lib.so.6`, will list details on glibc. Look in the listing for either 'nptl' or 'linuxthreads'. On debian systems, lib.so.6 is not executable but you can make it so. You can also do the following to determine library versions and which threading you are using: `% getconf GNU_LIBC_VERSION` and `% getconf GNU_LIBPTHREAD_VERSION`.

See [1086554] glibc 2.3.2 NPTL hang (Was bdbfrontier stall in...) [http://sourceforge.net/tracker/index.php?func=detail&aid=1086554&group_id=73833&atid=539099] for more on the issue.

13.1.2. [1093962] SSL handshake fails when server requests switch to SSL V3 [http://sourceforge.net/tracker/index.php?func=detail&aid=1093962&group_id=73833&atid=539099]

When connecting to a secure server, if the server wants to switch from SSL V2 to SSL V3 when client is using a SUN JVM, the connection fails. See issue 1093962for more.

13.1.3. Using old jobs or profiles with 1.4

You'll need to make one change to make your old order.xml files and profiles to run with Heritrix 1.4.x. Below is a diff that shows the change that needs to be made (The type of the `path` changed from `string` to `stringList`):

```
+++ order.xml   2005-02-01 13:12:34.000000000 -0800
@@ -162,7 +162,9 @@
         <string name="prefix">BT</string>
         <string name="suffix"></string>
         <integer name="max-size-bytes">100000000</integer>
-        <string name="path">arcs</string>
+        <stringList name="path">
+          <string>arcs</string>
+        </stringList>
         <integer name="pool-max-active">5</integer>
         <integer name="pool-max-wait">300000</integer>
       </newObject>
```

13.1.4. [1119644] frontier ConcurrentModificationException [https://sourceforge.net/tracker/?func=detail&atid=539099&aid=1119644&group_id=73833]

Sometimes you'll get a ConcurrentModificationException exception when you go to view or refresh the Frontier's report page. Workaround is to retry. The page should eventually come up.

13.1.5. New ARC file suffix

Pre-release 1.2.0, currently open ARC files that are being written to by the crawler were differentiated by an '.open' suffix. When the crawler finished writing, the suffix was removed. A new suffix has been introduced -- '.invalid' -- which the crawler will use to mark ARC files it thinks suspect -- usually because there was an IOException thrown during the writing of an ARC Record. Such ARCs need to be checked for validity. Run `% gzip -t` and `% ARCReader --strict` against all files with an '.invalid' suffix -- and any unclosed '.open' files present after a crawl has ended -- to check for corruption.

13.1.6. DNS lookups fail (-6 in crawl.log)

[1149470] all DNS attempts fail -6 [https://sourceforge.net/tracker/index.php?func=detail&aid=1149470&group_id=73833&atid=539099] discusses badly-formatted DNS records returned on windows platform that Heritrix fails to parse and it includes a pointer to a mailing list discussion of failed lookups on non-english windows. The issue includes description of a workaround.

13.1.7. FatalConfigurationException creating new job based on old

Older SUN JVMS -- pre-beta3 versions of the SUN JVM 1.5.0 for instance -- had an issue using nio copying files. Try upgrading your JVM. See [1178102] FCE on creation of new job based on job w/ overrides [http://sourceforge.net/tracker/index.php?func=detail&aid=1178102&group_id=73833&atid=539099] for more on this.

13.1.8. OutOfMemoryErrors (OOMEs)

Unusual pages -- pages of unorthodox structure, pages that contain thousands upon thousands of links -- will on occasion produce OOMEs.

There have been improvements regards memory usage running multiple jobs in series, Section 14.1.3, “Running more than one job in series throws OOME”, but starting up a new job after a long-running job can prompt OOMEs. Workaround for now is to restart Heritrix between the running of big jobs.

13.2. Changes

13.2.1. Berkeley DB Based Frontier

The BdbFrontier -- a frontier that keeps its queues of URIs in Berkeley DB Java Edition [http://www.sleepycat.com/products/je.shtml] databases -- has been made the default Frontier. Other core datastructures such as the queue of 'alreadyseen' URIs have also been moved into bdbje databases.

13.2.2. The IP in dns ARC Records

Dns entries in ARCs look like this:

```
dns:www.archive.org 207.241.238.254 20050310233154 text/dns 58 20050310233154
www.archive.org.        1600    IN      A       207.241.224.241
```

The above record is for the lookup of www.archive.org.

Previous to 1.4.0, the IP used on the ARC Record metaline -- the first line of an ARC Record entry (207.241.238.254 in the above example) -- was the IP of the host looked up. As of 1.4.0, we write the IP of the dns server that returned us the address looked up. Previous to this there was no recording of the dnsserver IP.

13.2.3. AdaptiveRevisitFrontier

A new, experimental Frontier with configurable revisiting policy and tools for noticing page change, etc.

13.2.4. DecidingScope and DecidingFilter

A.K.A New Scoping Model

A new, experimental scope and filter that allow the user to pick and choose from an assortment of ready-made decision rules and have each rule applied in an orderable sequence. The last non-PASS decision stands as the aggregate decision for the decide rule sequence.

13.2.5. Crawl Size Upper Bounds Update

Memory usage has been improved in this release. Previously RAM-based datastructures that grew without bound now are disk-backed kept in berkeley db databases. Where previous, see Section 17.1.1, “Crawl Size Upper Bounds”, Heritrix was unsuited for broad crawling, while still experimental, using default memory settings -- a heap of 256m -- broad-crawls of 5 to 6 days before encountering OutOfMemoryErrors (OOMEs) are now possible; longer if more heap is assigned. Where 10k hosts was an upper bound on narrow domain- or host-scoped crawls, now, using the default heap size, it should

now be possible to do 500k+ hosts.

Long-running crawls that encounter hundreds-of-thousands of hosts over the life of a crawl, or crawls started with hundreds-of-thousands of seeds, continue to throw OutOfMemoryErrors because there are still a few RAM-based datastructures that grow without bound left in Heritrix; the lists of queue names and internal structures inside 3rd party libraries used by Heritrix. These last few items we intend to address in a later release.

13.2.6. IBM JVM Redux

Testing with `IBM JVM 1.4.2 (Classic VM (build 1.4.2, J2RE 1.4.2 IBM build cxia32142sr1a-20050209 (JIT enabled: jitc))` using Heritrix 1.4.0, the SSL problem described in Section 14.1.1, “IBM JVM” is no longer present (All of our crawling of the last couple of months has been done on the latest SUN 1.5.0 JVMs).

Table 6. Changes

ID	Type	Summary	Open Date	By	Filer
958061 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=958061]	Add	[Post 1.0] New scoping model	2004-05-21	gojomo	gojomo
1165205 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1165205]	Add	Add links to issue tracking/ RFE to Heritrix' webapp	2005-03-17	nobody	ck-heritrix
1119580 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1119580]	Add	Integrate revisiting frontier	2005-02-09	kristinn_sig	stack-sf
1093609 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1093609]	Add	One-click recover	2004-12-30	gojomo	gojomo
1078008 [http://sourceforge.net/tracker/index.php?func=d	Add	Enable crawl-end at target compressed-ARC-data size	2004-12-02	stack-sf	gojomo

ID	Type	Summary	Open Date	By	Filer
etail&group_id=73833&atid=539102&aid=1078008]					
934577 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=934577]	Add	Need 'delete profile' option (like delete job)	2004-04-13	kristinn_sig	gojomo
1058302 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1058302]	Add	A 'dat' maker; A script to dump links	2004-11-01	stack-sf	stack-sf
1114133 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1114133]	Add	Add referer header	2005-02-01	stack-sf	stack-sf
1143892 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1143892]	Add	[contribution] SingleConnectionManager, range and close hdrs	2005-02-18	stack-sf	stack-sf
1055766 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1055766]	Add	Dates in logs are unreadable.	2004-10-27	gojomo	stack-sf
1111656 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1111656]	Add	Extractors should not extract if links already extracted	2005-01-28	stack-sf	stack-sf
1047437 [http://sourcefor	Add	Pause and alert on low-disk	2004-10-14	gojomo	gojomo

ID	Type	Summary	Open Date	By	Filer
ge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1047437]		conditions			
1104916 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1104916]	Add	Add info to candidateURI before scheduling	2005-01-18	stack-sf	stack-sf
953994 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=953994]	Add	Change arc download dir mid-crawl	2004-05-14	stack-sf	stack-sf
894467 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=894467]	Add	Stopping, pausing, checkpointing from command line/ scripts	2004-02-10	stack-sf	stack-sf
1096737 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1096737]	Add	[jmx] client pword and always start jmx server	2005-01-05	nobody	stack-sf
1090663 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1090663]	Add	Move BDB to core of Heritrix	2004-12-23	stack-sf	stack-sf
1092769 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1092769]	Add	[ARCReader] If garbage on end of record, report and skip it	2004-12-29	stack-sf	stack-sf

ID	Type	Summary	Open Date	By	Filer
1078016 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1078016]	Add	'Economic' frontier which defers low-value URIs	2004-12-02	gojomo	gojomo
1002704 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1002704]	Add	Evaluate Berkeley DB Frontier	2004-08-03	gojomo	stack-sf
1083315 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1083315]	Add	Update commons-pool, commons-collections, itext jars	2004-12-10	nobody	stack-sf
988276 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=988276]	Add	ARC writer pool config. to write multiple disks	2004-07-09	stack-sf	stack-sf
1078714 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1078714]	Add	Command-line insertion of URLs	2004-12-03	stack-sf	stack-sf
1069105 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1069105]	Add	Make auto seed add on redirect optional (if happens at all)	2004-11-18	gojomo	gojomo
1002707 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=100	Add	Fix heritrix shutdown (From Luca)	2004-08-03	nobody	stack-sf

ID	Type	Summary	Open Date	By	Filer
2707]					
1065736 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1065736]	Add	Recovery should optionally retain failures ('Ff')	2004-11-13	gojomo	gojomo
1057064 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1057064]	Add	HTTPRecorder's default buffer sizes should be configurable	2004-10-29	gojomo	gojomo
1045817 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1045817]	Add	Untangle heritrix from jetty	2004-10-12	stack-sf	stack-sf
1036720 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1036720]	Fix	NPE in ArcWriterProcessor.writeDns()	2004-09-28	stack-sf	gojomo
1178927 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1178927]	Fix	'submodules' map-edits not working for overrides/refinements	2005-04-07	gojomo	gojomo
1179530 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1179530]	Fix	NPE in FastBufferedOutputStream.close	2005-04-08	nobody	stack-sf
1184102 [http://sourceforge.net/tracker/index.php?func=detail&group_id	Fix	Frontier queues total still goes minus	2005-04-15	nobody	stack-sf

ID	Type	Summary	Open Date	By	Filer
=73833&atid=539099&aid=1184102]					
1179527 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1179527]	Fix	ARCWriter AsynchronousCloseException	2005-04-08	nobody	stack-sf
1096855 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1096855]	Fix	CME adding filters while crawling	2005-01-05	nobody	stack-sf
1080378 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1080378]	Fix	job config: settings 'remove'-component-then-submit lost job	2004-12-06	nobody	gojomo
1176788 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1176788]	Fix	hosts-report.txt is empty	2005-04-04	stack-sf	danavery
1172183 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1172183]	Fix	Delete URIs from frontier broken (CachedBdbBigMap.values()?)	2005-03-28	gojomo	gojomo
1178102 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1178102]	Fix	FCE on creation of new job based on job w/ overrides	2005-04-06	nobody	stack-sf
1178103 [http://sourceforge.net/tracker/in	Fix	hung bdb (12115 redux)	2005-04-06	nobody	stack-sf

ID	Type	Summary	Open Date	By	Filer
dex.php?func=detail&group_id=73833&atid=539099&aid=1178103]					
1169459 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1169459]	Fix	CachedBdbBigMap double-close in finialize()	2005-03-23	gojomo	gojomo
1177462 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1177462]	Fix	RIS#readFullyOrUntil IOE/ timeout	2005-04-05	nobody	stack-sf
1149470 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1149470]	Fix	all DNS attempts fail -6	2005-02-22	nobody	jsleeman
1156363 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1156363]	Fix	Flash SWF Extractor Unexpected end of input	2005-03-03	nobody	stack-sf
1170562 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1170562]	Fix	npe in extractorjs doing broad crawl w/ HEAD	2005-03-25	nobody	stack-sf
1121567 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1121567]	Fix	Heritrix 1.3.0 crashes hard (JVM SIGSEV)	2005-02-12	nobody	stack-sf
1103015	Fix	If filter in main	2005-01-15	nobody	frodobay

ID	Type	Summary	Open Date	By	Filer
[http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1103015]		scope disabled heritrix aborts imme			
1054219 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1054219]	Fix	Links not extracted from mislabelled (text/plain) MIME type	2004-10-25	gojomo	gojomo
1024120 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1024120]	Fix	Lost crawl job after terminate running job with jobs pending	2004-09-07	stack-sf	stack-sf
1078094 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1078094]	Fix	www-strip canonicalization unintended exclusion of redirect	2004-12-02	stack-sf	gojomo
1157085 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1157085]	Fix	DNS records in ARCs should use DNS server IP	2005-03-04	stack-sf	gojomo
1157385 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1157385]	Fix	Crawler not making progress -- thread deadlock	2005-03-05	stack-sf	ia_igor
1158270 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1158270]	Fix	isMultibyteEncoding: Uncaught UnsupportedOperationException	2005-03-07	stack-sf	ck-heritrix

ID	Type	Summary	Open Date	By	Filer
1080925 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1080925]	Fix	MultiThreaded-Connection-Manager bottle-neck	2004-12-07	stack-sf	gojomo
1157372 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1157372]	Fix	missing space in progress-statistics.log	2005-03-05	stack-sf	ia_igor
1153927 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1153927]	Fix	npe in Extractor-HTML#innerProcess	2005-02-28	stack-sf	stack-sf
1155641 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1155641]	Fix	"Illegal response body offset" in ReplayCharSequenceFactory	2005-03-02	stack-sf	gojomo
1154673 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1154673]	Fix	ensure IPs match from DNS, used in HTTP, logged in ARC	2005-03-01	stack-sf	gojomo
1002138 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1002138]	Fix	swf extractor flash lib prints glyphcount on stdout	2004-08-02	nobody	stack-sf
1077924 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=107	Fix	crawl.log timestamps out-of-order	2004-12-02	gojomo	gojomo

ID	Type	Summary	Open Date	By	Filer
7924]					
1066573 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1066573]	Fix	sometimes job based-on other job uses older job name	2004-11-15	gojomo	gojomo
1102755 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1102755]	Fix	seeds text area truncates seeds; big seed lists break config	2005-01-14	gojomo	gojomo
1002164 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1002164]	Fix	OOM hit very early broad-crawling	2004-08-02	stack-sf	gojomo
1006970 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1006970]	Fix	UI list-ordering inconsistent	2004-08-10	gojomo	gojomo
1092937 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1092937]	Fix	UI/Settings - Expert Toggle loses user data	2004-12-29	nobody	nobody
1152358 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1152358]	Fix	OOM in postselector	2005-02-26	nobody	orion2598
1068403 [http://sourceforge.net/tracker/index.php?func=detail&group_id	Fix	ARCWriter gzip deflate hang	2004-11-17	nobody	stack-sf

ID	Type	Summary	Open Date	By	Filer
=73833&atid=539099&aid=1068403]					
1123906 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1123906]	Fix	ARCWriter alerts if Content-Type is null	2005-02-16	nobody	ck-heritrix
1124029 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1124029]	Fix	Bad synchronization causes NPE in StatisticsTracker	2005-02-16	nobody	ck-heritrix
1055789 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1055789]	Fix	ARCWriter 'Gap' errors should be more prominent	2004-10-27	stack-sf	gojomo
1123859 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1123859]	Fix	Change in ExtractorHTML triggers NullPointerExceptions	2005-02-16	nobody	ck-heritrix
1093073 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1093073]	Fix	StackOverflowError shouldn't kill crawl	2004-12-29	gojomo	gojomo
1068370 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1068370]	Fix	[Flash] OOMEs on a particular URL	2004-11-17	nobody	stack-sf
1108153 [http://sourceforge.net/tracker/in	Fix	unwritable ARCs directory barely notice-	2005-01-23	nobody	gojomo

ID	Type	Summary	Open Date	By	Filer
dex.php?func=detail&group_id=73833&atid=539099&aid=1108153]		able			
1023929 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1023929]	Fix	"&amp" converted to "&" in preselector override regex	2004-09-07	gojomo	danavery:
1083428 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1083428]	Fix	remove profile function in WUI?	2004-12-11	nobody	zhousp
1068384 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1068384]	Fix	deleting all(?) from queue corrupts frontier, kills crawl	2004-11-17	gojomo	gojomo
1106469 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1106469]	Fix	ExtractorCSS regexp taking 'forever' on small document	2005-01-20	gojomo	gojomo
1116204 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1116204]	Fix	FetchDNS doesn't work (bug in dnsjava)	2005-02-04	nobody	nobody
1103838 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1103838]	Fix	Redirect problem (Stops crawling after 3)	2005-01-17	nobody	nobody
1119686	Fix	oversight in	2005-02-09	nobody	frodobay

ID	Type	Summary	Open Date	By	Filer
[http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1119686]		CrawlURI; missing check for null			
1060508 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1060508]	Fix	[uuri] port StringIndex-OutOfBound-sExceptionn	2004-11-04	stack-sf	stack-sf
1101831 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1101831]	Fix	NPE in ROS#record	2005-01-13	nobody	stack-sf
1114285 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1114285]	Fix	Old profile/jobs won't work with HEAD (1.4)	2005-02-01	nobody	stack-sf
1062621 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1062621]	Fix	First arc record length is off by one	2004-11-08	stack-sf	stack-sf
1117916 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1117916]	Fix	PDFParser URL extraction bug	2005-02-07	nobody	benlitchfield
1113977 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1113977]	Fix	User Agent is tolowercased	2005-02-01	nobody	nobody

ID	Type	Summary	Open Date	By	Filer
1113470 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1113470]	Fix	Exception in Modules Tab	2005-01-31	nobody	nobody
1109521 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1109521]	Fix	Hung Thread in StatisticsTracker	2005-01-25	stack-sf	ia_igor
1107304 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1107304]	Fix	Failed create new job based on job with absolute settings	2005-01-22	nobody	frodobay
1000865 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1000865]	Fix	Long random pauses where no progress is made	2004-07-30	nobody	gojomo
1095952 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1095952]	Fix	InvalidJobFileException: Status .. 'RUNNING'	2005-01-04	nobody	stack-sf
1095453 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1095453]	Fix	heritrix wont start with fedora core 3	2005-01-03	nobody	nobody
1092135 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=109	Fix	crawl.log hashes wrong for captures > 64K	2004-12-28	gojomo	gojomo

ID	Type	Summary	Open Date	By	Filer
2135]					
1103133 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1103133]	Fix	deadlock in ip-politeness re-queueing	2005-01-15	gojomo	gojomo
1102771 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1102771]	Fix	SURTs-from-seeds may lack trailing comma	2005-01-14	gojomo	gojomo
1101396 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1101396]	Fix	JS extr. does not parse spec. links starting w/ ./ or ../	2005-01-12	nobody	ia_igor
1100658 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1100658]	Fix	update to [1100467] maven 1.0.2 build problem	2005-01-11	stack-sf	nobody
1101138 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1101138]	Fix	Update ant and httpclient jars	2005-01-12	stack-sf	stack-sf
1098217 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1098217]	Fix	ReplayCharSequence.toString() is broken	2005-01-07	nobody	stack-sf
1093627 [http://sourceforge.net/tracker/index.php?func=detail&group_id	Fix	[robots] robots.txt mid-fetch aborted gives open access	2004-12-30	nobody	stack-sf

ID	Type	Summary	Open Date	By	Filer
=73833&atid=539099&aid=1093627]					
1093614 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1093614]	Fix	midfetch abort doesn't	2004-12-30	nobody	stack-sf
1082358 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1082358]	Fix	[uuri] String index out of range: 0	2004-12-09	stack-sf	stack-sf
1086554 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1086554]	Fix	glibc 2.3.2 NPTL hang (Was bdbfrontier stall in...)	2004-12-16	stack-sf	stack-sf
1072035 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1072035]	Fix	[uuri] Underscore in host messes up port parsing	2004-11-23	stack-sf	stack-sf
1043251 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1043251]	Fix	better/longer dns retries on lookup failure	2004-10-08	gojomo	stack-sf
1090911 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1090911]	Fix	NPE in ServerCache	2004-12-24	stack-sf	stack-sf
1080926 [http://sourceforge.net/tracker/in	Fix	reducing max-toe-threads has no effect	2004-12-07	gojomo	gojomo

ID	Type	Summary	Open Date	By	Filer
dex.php?func=detail&group_id=73833&atid=539099&aid=1080926]					
1088788 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1088788]	Fix	NPE in TextUtils.freeMatcher()	2004-12-20	stack-sf	gojomo
1082570 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1082570]	Fix	heritrix.log ignored	2004-12-09	nobody	stack-sf
1078503 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1078503]	Fix	Edit configuration in UI gives NPE	2004-12-03	nobody	stack-sf
1055592 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1055592]	Fix	terminated crawl still hogging memory, causing OOM	2004-10-27	nobody	gojomo
1081770 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1081770]	Fix	quick-override accepts domain w/spaces, lost checkboxes	2004-12-08	gojomo	gojomo
1080827 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1080827]	Fix	Browser hangs when hundreds of seeds	2004-12-07	nobody	stack-sf
1047396	Fix	OOM in Bdb-	2004-10-14	nobody	gojomo

ID	Type	Summary	Open Date	By	Filer
[http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1047396]		Frontier/ nio.Bits -- with plenty of heap left			
1078581 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1078581]	Fix	DomainSensitiveFrontier never finishes	2004-12-03	nobody	stack-sf
1076251 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1076251]	Fix	Upgrade bdbje 1.7.0 (WAS: Checkpointer thread ...)	2004-11-30	nobody	stack-sf
1072192 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1072192]	Fix	bdbfrontier No locks available	2004-11-23	nobody	stack-sf
1031499 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1031499]	Fix	Deleted pending jobs show as pending in restart.	2004-09-20	stack-sf	stack-sf

14. Release 1.2.0 - 2004-11-16

Added IP-based politeness, configurable URI-canonicalization, and mid-fetch abort. Lots of Bug fixes.

14.1. Known Limitations

14.1.1. IBM JVM

The IBM JVM generally is more performant than SUN JVMs. It also emits more detailed heap dumps. That said, new Heritrix 1.2.0 features may not work on the IBM JVM.

14.1.1.1. HTTPS

Heritrix 1.2.0 uses the new HttpClient 3.0x library which allows the setting of socket read timeouts. Connections to https sites fail if using the IBM JVM.

The IBM JVM 141 (cxia321411-20030930) NPEs setting the NoTcpDelay.

```
java.lang.NullPointerException
   at com.ibm.jsse.bf.setTcpNoDelay(Unknown Source)
   at org.apache.commons.httpclient.HttpConnection.open(HttpConnection.java:683)
   at org.apache.commons.httpclient.MultiThreadedHttpConnectionManager$HttpConnect
```

Using the IBM JVM 142, its saying SSL connection not open when we go to use inputstreams:

```
java.net.SocketException: Socket is not connected
   at java.net.Socket.getInputStream(Socket.java:726)     at com.ibm.jsse.bs.getIn
   at org.apache.commons.httpclient.HttpConnection.open(HttpConnection.java:715)
   at org.apache.commons.httpclient.MultiThreadedHttpConnectionManager$HttpConnect
```

Newer versions of the httpclient library may address this (Current version is alpha2).

14.1.2. Jobs don't show in UI when a bunch are run back-to-back

If more than one job waiting in the queue of pending jobs, then the second job often won't show in the UI; The UI says its running but its not possible to see a status bar on the running job. See [1024120] Lost crawl job after terminate running job with jobs pending [https://sourceforge.net/tracker/?func=detail&aid=1024120&group_id=73833&atid=539099]. For now, the workaround is to study the running job by viewing the crawl job logs on disk (Oddly, the 3rd queued up job will start to show in the UI again).

14.1.3. Running more than one job in series throws OOME

OutOfMemoryExceptions are frequent when jobs are run in series. [1055592] terminated crawl still hogging memory, causing OOM [https://sourceforge.net/tracker/index.php?func=detail&aid=1055592&group_id=73833&atid=539099]. For now, restart Heritrix between the running of jobs.

14.2. Changes

Table 7. Changes

ID	Type	Summary	Open Date	By
1067095 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1067095]	Add	Hang in http fetcher when mid-fetch aborts	2004-11-15	stack-sf
1066804 [http://sourceforge.net/track-	Add	Allow specification of heritrix_out.log filename	2004-11-15	stack-sf

ID	Type	Summary	Open Date	By
er/in-dex.php?func=detail&group_id=73833&atid=539102&aid=1066804]				
903845 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539102&aid=903845]	Add	IP-based politeness	2004-10-28	gojomo
1054849 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539102&aid=1054849]	Add	Recover from crawl initialized with a recovery log	2004-10-26	stack-sf
1054851 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539102&aid=1054851]	Add	Import gzipped or non-gzipped recov-ery log	2004-10-26	stack-sf
1050378 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539102&aid=1050378]	Add	Add bdb alreadyseen option to hostsqueuesfron-tier	2004-10-19	stack-sf
973881 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539102&aid=973881]	Add	Force generation of report files	2004-06-16	stack-sf
1010883 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539102&aid=1010883]	Add	Scripts to generate end-of-job reports	2004-08-17	danavery

ID	Type	Summary	Open Date	By
988277 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=988277]	Add	[Need feedback] "Done with ARC file" event	2004-07-09	stack-sf
1044977 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1044977]	Add	Logging of scope-rejected URIs	2004-10-11	stack-sf
902970 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=902970]	Add	HTTPClient should use supplied IP / avoid DNS lookup	2004-02-23	stack-sf
903093 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=903093]	Add	Setting of Integer.MAX_VALUE is ugly	2004-02-23	stack-sf
900004 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=900004]	Add	canonicalization of URIs for alreadyIncluded testing	2004-02-18	stack-sf
941072 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=941072]	Add	Allow operator-configured mid-HTTP-fetch filters	2004-04-23	stack-sf
1037891 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833	Add	Cmdline defaults in properties file	2004-09-30	stack-sf

ID	Type	Summary	Open Date	By
&atid=539102&aid=1037891]				
1037304 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1037304]	Add	Upgrade httpclient to 3.0.x	2004-09-29	stack-sf
994141 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=994141]	Add	Update build to use maven 1.0	2004-07-19	stack-sf
1002336 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=1002336]	Add	Figure what profiler to use	2004-08-02	stack-sf
1064887 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1064887]	Fix	http and https prerequisites contention	2004-11-11	stack-sf
1062604 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1062604]	Fix	Seed to SURT coversion issuesI	22004-11-08	stack-sf
11061795 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=11061795]	Fix	ServerCache HashMaps access thread-safetyI	22004-11-06	gojomo
1060589 [http://sourceforge.net/track-	Fix	Can't open logs of old jobs post-restart in UII	22004-11-04	stack-sf

ID	Type	Summary	Open Date	By
er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=1060589]				
1058565 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=1058565]	Fix	Non-default 'logs' location doesn't show in web UI	2004-11-010	stack-sf
1058568 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=1058568]	Fix	IMG 'lowsrc' may not be extracted	2004-11-010	stack-sf
1055854 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=1055854]	Fix	completed crawls show as 'aborted by user'	2004-10-270	gojomo
1059237 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=1059237]	Fix	MultiThreadedHt-tpConnectionMan-ager https already connected	2004-11-020	stack-sf
1052578 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=1052578]	Fix	recovery log of re-covered crawl in-sufficient to recov-er	2004-10-220	stack-sf
908690 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=908690]	Fix	Some dates are GMT, others are not *	2004-03-020	gojomo

ID	Type	Summary	Open Date	By
958096 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=958096]	Fix	Flushing CrawlServers problematic *	2004-05-210	gojomo
1052570 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1052570]	Fix	Threads contend for scratch files (after kill/readFully/Gap)	2004-10-220	gojomo
1033701 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1033701]	Fix	incorrect number of total active threads *	2004-09-230	gojomo
1000840 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1000840]	Fix	diskincludedfrontier performance is awful	2004-07-30	gojomo
1043251 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1043251]	Fix	better/longer dns retries on lookup failure	2004-10-08	gojomo
1051072 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1051072]	Fix	ExtractorHTML takes forever on worst-case HTML	2004-10-20	gojomo
1051916 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833	Fix	ExtractorJS takes forever on worst-case JS	2004-10-21	gojomo

ID	Type	Summary	Open Date	By
&atid=539099&aid=1051916]				
1050238 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1050238]	Fix	jdk required (doc implies jre)	2004-10-19	stack-sf
1038135 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1038135]	Fix	prerequisite hysteresis/robots ahead of dns	2004-09-30	gojomo
1015728 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1015728]	Fix	Crawl upper time/size bounds ignored	2004-08-24	gojomo
1002356 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1002356]	Fix	timing issue on crawl-start & runtime stat	2004-08-02	gojomo
1002332 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1002332]	Fix	inactiveQueuesMemoryLoadTarget mechanism behaves poorly	2004-08-02	gojomo
1045016 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1045016]	Fix	DNS URIs don't get override settings	2004-10-11	gojomo
998184 [http://sourceforge.net/track-	Fix	Gzipped recover log corrupt at end; last < 32K unre-	2004-07-26	gojomo

ID	Type	Summary	Open Date	By
er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=998184]		coverable		
998272 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=998272]	Fix	No crawl if host-queues-memory-capacity = 0	2004-07-26	stack-sf
1002335 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=1002335]	Fix	frontier report un-usable in big crawls; frontier info needed	2004-08-02	gojomo
984390 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=984390]	Fix	Build fails: "rws" mode and Mac OS X interact badly	2004-07-02	stack-sf
1000929 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=1000929]	Fix	fatal runtimeexcep-tions in frontier give no info in web UI	2004-07-30	gojomo
964625 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=964625]	Fix	seed parser *too* lenient	2004-06-01	johnerik
980051 [http://sourceforge.net/track-er/in-dex.php?func=detail&group_id=73833&atid=539099&aid=980051]	Fix	Auth unsupported logged to console	2004-06-25	stack-sf

ID	Type	Summary	Open Date	By
1002146 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1002146]	Fix	bad queue keys: shouldn't be URIs; should be handled better	2004-08-02	stack-sf
1046696 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1046696]	Fix	UURIFactory.validateEscaping() -> IllegalArgumentException	2004-10-13	stack-sf
1045736 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1045736]	Fix	ARCReader crashes if zero-length gzip record	2004-10-12	stack-sf
1002144 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1002144]	Fix	[UURI] Catch bad-encoding earlier	2004-08-02	stack-sf
1036680 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1036680]	Fix	PathDepthFilter innerAccepts SEVERE log: "Failed getPath..."	2004-09-28	stack-sf
1045847 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1045847]	Fix	Unnecessary toString() in ExtractorHTML.processScriptCode()	2004-10-12	gojomo
1044527 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833	Fix	Domain names in 'overrides' are not in alphabetical order	2004-10-11	kristinn_sig

ID	Type	Summary	Open Date	By
&atid=539099&aid=1044527]				
1012639 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1012639]	Fix	If CC timesout selftest, no build failed message	2004-08-19	stack-sf
1012642 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1012642]	Fix	selftest hanging because no crawl stop event	2004-08-19	stack-sf
931565 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=931565]	Fix	CrawlStateUpdater - NullPointerException	2004-04-08	stack-sf
973294 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=973294]	Fix	NoSuchElementException in URI queues halts crawling	2004-06-15	gojomo
1033657 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1033657]	Fix	[UURI] >2047 AFTER escaping (Stops crawl)	2004-09-23	stack-sf
1010966 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1010966]	Fix	crawl.log has URIs with spaces in them	2004-08-17	stack-sf
963970 [http://sourceforge.net/track-	Fix	unfetchable URI schemes should never be queued	2004-05-31	gojomo

ID	Type	Summary	Open Date	By
er/index.php?func=detail&group_id=73833&atid=539099&aid=963970]				
1031607 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1031607]	Fix	KeyedQueue server<->key mismatch noted: pf-buser<->mprsrv.agr	2004-09-20	stack-sf
1031525 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1031525]	Fix	NPE reading override	2004-09-20	stack-sf
1031168 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1031168]	Fix	Wrong handling of date in ARCRecordMetaData	2004-09-20	johnerik

15. Release 1.0.4 - 2004-09-22

Bug fix.

15.1. Changes

Table 8. Changes

ID	Type	Summary	Open Date
1010966 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1010966]	Fix	crawl.log has URIs with spaces in them	2004-08-17

16. Release 1.0.2 - 2004-09-14

Bug fixes.

16.1. Changes

Table 9. Changes

ID	Type	Summary
1020770 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1020770]	Fix	old crawls stick around, consuming memory
1002319 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1002319]	Fix	Terminating paused crawl leaves zombie threads
935146 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=935146]	Fix	Excessive ARCWriterPool timeouts
1010859 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1010859]	Fix	Per host overrides not taking effect.
1014732 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1014732]	Fix	document size limit not working
1012520 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1012520]	Fix	UURI.length() > 2k
1010966 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=1010966]	Fix	crawl.log has URIs with spaces in them stack-sf

17. Release 1.0.0 - 2004-08-06

Added new prefix ('SURT') scope and filter, compression of recovery log, mass adding of URIs to running crawler, crawling via a http proxy, adding of headers to request, improved out-of-the-box defaults, hash of content to crawl log and to arcreader output, and many bug fixes.

17.1. Known Limitations

17.1.1. Crawl Size Upper Bounds

Heritrix 1.0.0 uses disk-based queues to hold any number of pending URIs bounded only by available disk space, but still relies on in-memory structures to efficiently track all discovered hosts and previously-scheduled URIs. Crawls whose total scheduled URIs or discovered hosts exhaust all available memory will trigger out-of-memory errors, which freeze a crawl at the point of the error.

With the default settings, and an assignment of a 256MB Java heap to the Heritrix process, crawling which discovers up to 10 000 hosts, and schedules over 6 000 000 URIs, should be possible. Discovery of higher numbers of URIs/hosts will likely trigger out-of-memory problems unless a larger java heap was assigned at startup.

Broad crawls -- those using the BroadScope or ranging over domains with many subdomains -- can easily and quickly exceed these parameters. Thus broad crawls in Heritrix 1.0.0 are not recommended, except for experimental purposes.

Narrower crawls, restricted to specific hosts or domains a limited number of subdomains, can run for a week or more, collecting millions of resources. Larger heaps can allow crawls to run into the tens of millions of collected URIS, and tens of thousands of discovered hosts.

An experimental alternate Frontier, the DiskIncludedFrontier, is also available via the 'Modules' crawl configuration tab. It uses a capped amount of memory plus disk storage to remember any number of scheduled URIs, but its performance is poor and it has not received the same testing as our default Frontier. The memory cost of additional discovered hosts continues to rise without limit when using a DiskIncludedFrontier.

Future versions of Heritrix will include other frontier implementations allowing larger and unbounded crawls with minimal performance penalties.

17.1.2. [958055] Seed ConcurrentModificationException [https://sourceforge.net/tracker/?func=detail&aid=958055&group_id=73833&atid=539099]

Its possible to get ConcurrentModificationsException editing options on a running crawl.

17.1.2.1. Workaround

Pause the crawl when making changes to crawl options.

17.1.3. [984390] Build fails: "rws" mode and Mac OS X interact badly [https://sourceforge.net/tracker/?func=detail&aid=984390&group_id=73833&atid=539099]

On macintoshes and linux kernel version 2.6, heritrix fails to build (unit tests fail).

17.1.3.1. Workaround

See issue, [984390] Build fails: "rws" mode and Mac OS X interact badly [https://sourceforge.net/tracker/?func=detail&aid=984390&group_id=73833&atid=539099], for source code workaround edit.

17.1.4. [955975] Build fails: JVM and kernel 2.6+ (Was 2 tests fail...) [https://sourceforge.net/tracker/index.php?func=detail&aid=955975&group_id=73833&atid=539099]

Heritrix fails to build on linux kernel 2.6.

17.1.4.1. Workaround

Build fails unless you use a JDK in advance of pedigree 1.5 beta 2 (It works with jdk1.5.0-rc). See [955975] Build fails: JVM and kernel 2.6+ (Was 2 tests fail...) [https://sourceforge.net/tracker/index.php?func=detail&aid=955975&group_id=73833&atid=539099] and above [#984390].

17.2. Changes

Table 10. Changes

ID	Type	Summary
939679 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=939679]	Add	Mass-add URIs to running crawl and force reconsideration
986977 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=986977]	Add	SurtPrefix scope (and filter)
989816 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=989816]	Add	Specification of default CharSequence charset
983001 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=983001]	Add	crawl.log entries all on one line
869584 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=869584]	Add	Hash content-bodies, show in logs (and future ARCs)
964581 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=964581]	Add	option to preference (quick-get) embeds
964493 [http://sourceforge.net/tracker/index.php?func=detail&group_id=7	Add	Compress recover.log

ID	Type	Summary
3833&atid=539102&aid=964493]		
988106 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=988106]	Add	[UURI] 'http:///...' converted to 'http://...'
926143 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=926143]	Add	enable use through HTTP proxy
945922 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=945922]	Add	Allow adding (subtracting?) http headers
983109 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=983109]	Add	Improved out-of-the-box defaults
982909 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=982909]	Add	ARCWriter makes FAT gzip header
925734 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539102&aid=925734]	Add	exponential backoff URI/host retries
-	Fix	Total data "written" isn't necessarily written (wording)
-	Fix	embeds within scope problem
-	Fix	NPE clearing alerts
-	Fix	arcmetadata repeated once for every domain config
-	Fix	CCE deserializing diskqueue [Was: IllegalArgumentExcepti...]
-	Fix	no docs for recovery-journal feature
-	Fix	Pause/Terminate ignored on 2.6 kernel 1.5 JVM
-	Fix	Investigate "Relative URI but no base"
-	Fix	User-Agent should be able to mimic Mozilla (as does Google)
-	Fix	referral URL should be stored in recover.log

ID	Type	Summary
-	Fix	ToeThreads hung in FetchDNS after Pause
-	Fix	robots.txt lookup for different ports on same host
-	Fix	Empty log percentages displayed as NaN%
-	Fix	UURI doubly-encodes %XX sequences
-	Fix	Single settings change causes two versions to be created
-	Fix	New IA debian image is 2.6 (Was: Build fails: JVM and ...)
-	Fix	NPE in PathDepthFilter
-	Fix	[investigate & rule out] Thread report deadlock risks
-	Fix	jetty susceptible to DoS attack
-	Fix	'ignore' robots does not ignore meta nofollow
-	Fix	URI Syntax Errors stop page parsing.
-	Fix	NPE in ExtractorHTML/TextUtils.getMatcher()
-	Fix	ARCReader: Failed to find GZIP MAGIC
-	Fix	javascript embedded URLs
-	Fix	NoClassDefFoundError when starting a job
-	Fix	Max number of deferrals hard-coded to 10.
-	Fix	Frontier report thread safety problems?
-	Fix	ARCReader hanging
-	Fix	log-browsing by regexp outofmemoryerror
-	Fix	Deferred URLs due the DNS problem -- Heritrix(-50)-Deferred
-	Fix	Assertion failures shouldn't be more fatal than Runtime Exc.
-	Fix	min-interval is superfluous; remove
-	Fix	crawl doesn't end when using valence > 1
-	Fix	Giant (in # of files) state directory problematic
-	Fix	robots-expiration units, default wrong

ID	Type	Summary
-	Fix	NoSuchElementException in URI queues halts crawling
-	Fix	#anchor links not trimmed, and thus recrawled
-	Fix	arc's filedesc file name includes .gz
-	Fix	[denmark-workshop] Cookie mangling
-	Fix	HttpException: Unable to parse header
-	Fix	bogus ARC-header when no Content-type
-	Fix	paths when crawling without UI
-	Fix	domain scope leakage

18. Release 0.10.0 - 2004-06-04

Release for second heritrix workshop, Copenhagen 06/2004 (1.0.0 first release candidate). Added site-first prioritization, fixed link extraction of multibyte URIs, added metadata to arcs as xml, changed arc naming template, new user and developer manuals, added basic/digest auth and http post/get login facility, and added help to UI. Bug fixes.

18.1. Changes

Table 11. Changes

ID	Type	Summary
896769 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896769]	Add	job report: show 'active' hosts, show more size totals
896772 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896772]	Add	"Site-first"/'frontline' prioritization
956614 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=956614]	Add	multiple open http connections per host needed
896674 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896674]	Add	Add help to web UI

ID	Type	Summary
964931 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=964931]	Add	When a host last had a completed URI shown in crawl report
958335 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=958335]	Add	Encode multibyte URIs using page charset before queuing
909246 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=909246]	Add	One src for site, help, and readme docs.
936684 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=936684]	Add	identifying ARCs: unique names, header records
930667 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=930667]	Add	Resetting arc file counter for every job.
863318 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=863318]	Add	ARCs need better headers
908507 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=908507]	Add	Specify location of jobs dir
914301 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=914301]	Add	Logging in (HTTP POST, Basic Auth, etc.)
944066 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=944066]	Add	Update dnsjava from 1.5 to 1.6.2 (Fix NPE)
966168 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=966168]	Fix	crawl.log entries without annotations end with a space
966172 [http://sourceforge.net/tracker/ind	Fix	An issue with arc names' date and serial number alignment

ID	Type	Summary
ex.php?func=detail&group_id=73833amp;atid=539099amp;aid=966172]		
957963 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=957963]	Fix	Output of warning message leads to NullPointerExceptions
963965 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=963965]	Fix	Either UURI or ExtractHTML should strip whitespace better
965267 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=965267]	Fix	Maximum documents not enforced
965308 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=965308]	Fix	NPE in path depth filter
934549 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=934549]	Fix	embed/speculative inclusion too loose
962899 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=962899]	Fix	UnsupportedCharsetException handled awkwardly
962892 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=962892]	Fix	UURI accepting/creating unUsable URIs (bad hosts)
860733 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=860733]	Fix	CachingDiskLongFPSet UI availability
954130 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=954130]	Fix	Crawls slow till change a setting
961867 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=9	Fix	zero link-hops should work

ID	Type	Summary
61867]		
942627 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=942627]	Fix	multiple robots.txt URLs in the "default" frontier
957941 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=957941]	Fix	NPE in ExtractorHTML#isHtmlExpectedHere
953718 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=953718]	Fix	Unwanted behavior with seed redirection
952636 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=952636]	Fix	Link extraction failing
863315 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=863315]	Fix	Memory issues: Frontier.snoozeQueue
903838 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=903838]	Fix	Transitive scope confusion, may not work as expected
955345 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=955345]	Fix	Wrong stats after deleting URIs from Frontier
952276 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=952276]	Fix	NoSuchElementException in admin/reports/frontier.jsp
952665 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=952665]	Fix	Alert: Authentication scheme(s) not supported
936702 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=936702]	Fix	IP validity: units, TTL vs. setting
951582	Fix	ConcurrentModificationExcep-

ID	Type	Summary
[http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=951582]		tion in DomainScope focus filter
949489 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=949489]	Fix	ConcurrentModificationException terminate job
949551 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=949551]	Fix	Authentication bug
948898 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=948898]	Fix	terminate running crawl == NPE
927940 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=927940]	Fix	java.net.URI parses %20 but getHost null
874220 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=874220]	Fix	NPE in java.net.URI.encode
808270 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=808270]	Fix	java.net.URI chokes on hosts_with_underscores
788277 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=788277]	Fix	Doing separate DNS lookup for same host
910120 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=910120]	Fix	java.net.URI#getHost fails when leading digit
949548 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=949548]	Fix	Constraining java URI class
943373 [http://sourceforge.net/tracker/index.php?func=detail&group_id=7	Fix	Same CrawlServer instance for http & https.

ID	Type	Summary
3833amp;atid=539099amp;aid=943373]		
887999 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=887999]	Fix	Broad crawl/ too many open files
926912 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=926912]	Fix	multiple charset headers + long lines
926338 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=926338]	Fix	Corrupted blue image in progress bars
896757 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=896757]	Fix	NPEs in Andy's Th-Fri Crawl + NPE in RIS
922080 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=922080]	Fix	IllegalArgumentEx/ReplayCharSequenceFactory (offset vs. size
935271 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=935271]	Fix	FTP URIs in seeds interpreted as HTTP
945923 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=945923]	Fix	maven rc2 won't make src distribution
947754 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=947754]	Fix	Corrupted arc files on termination of job
931269 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=931269]	Fix	https exception: java.io.IOException: SSL failure
935146 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833amp;atid=539099amp;aid=935146]	Fix	Excessive ARCWriterPool timeouts:

19. Release 0.10.0 - 2004-06-04

Fixes to build with maven rc2+.

19.1. Changes

Table 12. Changes

ID	Type	Summary
962361 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=962361]	Fix	080 doesn't build with maven rc2+

20. Release 0.8.0 - 2004-05-18

Release (and branch heritrix-0_8 made at the heritrix-0_7_1 tag) because of concurrentmodificationexceptions if tens of seeds supplied and to fix domain-scope leakage. Also, made continuous build publically available, incorporated integration selftest into build, made it a maven-build only (ant-build no longer supported), added day/night configurations (refinements), ameliorated too-many-open files, added exploit of http-header content-type charset creating character streams, and heritrix now crawls ssl sites. UI improvements include red start by bad configuration, precompilation, and delineation of advanced settings. Many bug fixes.

20.1. Synopsis

20.2. Changes

Table 13. Changes

ID	Type	Summary
939032 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=939032]	Add	integrate selftest into cruisecontrol build
903078 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=903078]	Add	On reedit, red star by bad attribute setting.
935215 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=935215]	Add	day/night configurations

ID	Type	Summary
928745 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=928745]	Add	UI should only write changed config
908723 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=908723]	Add	record of settings changes should be kept
909249 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=909249]	Add	Only one build, not two
925614 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=925614]	Add	maven-only build rather than ant & maven
877295 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=877295]	Add	ARCWriter should use a pool of open files -- if it helps
899226 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=899226]	Add	Precompile UI pages
896798 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896798]	Add	UI should be split into common/uncommon settings
895341 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=895341]	Add	UI web pages need to be more responsive
955527 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=955527]	Fix	domain scope leakage
943770 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=943770]	Fix	ConcurrentModificationExceptions
943768 [http://sourceforge.net/tracker/ind	Fix	Too many open files

ID	Type	Summary
ex.php?func=detail&group_id=73833&atid=539099&aid=943768]		
943781 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=943781]	Fix	ConcurrentModificationExceptions
943453 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=943453]	Fix	empty seeds-report.txt
903092 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=903092]	Fix	Doc. assumes bash. Allow tcsh/csh
908419 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=908419]	Fix	script heritrix.sh goes into infinite loop
922104 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=922104]	Fix	heritrix.sh launch file path weirdness
935122 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=935122]	Fix	ToeThreads hung in ExtractorHTML after Pause
938591 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=938591]	Fix	IllegalCharsetNameException: Windows-1256
934642 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=934642]	Fix	No doc-files/package.html in javadoc.
815544 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=815544]	Fix	embed-count sensitivity WRT redirects, preconditions
936610 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=936610	Fix	Refinement limits are not always saved

ID	Type	Summary
]		
935340 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=935340]	Fix	NPE exception in getMBeanInfo(settings)
904767 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=904767]	Fix	Untried CrawlURIs should have clear status code
914287 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=914287]	Fix	Thread underutilization in broad crawls
930736 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=930736]	Fix	KeyedQueue showing EMPTY status, but the length is 1.
934585 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=934585]	Fix	NPE in XMLSettingsHandler.recursiveFindFiles()
935352 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=935352]	Fix	Failed DNS does not have intended impact
896764 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896764]	Fix	ftp URIs are retried
848661 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=848661]	Fix	Refetching of robots and/or DNS broken
935221 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=935221]	Fix	NPE switching to 'expert' settings in HEAD
896779 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896779]	Fix	rss extractor
896775	Fix	JS extractor clueless on relative

ID	Type	Summary
[http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896775]		URIs
913214 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=913214]	Fix	converting URI's '\' into '/' character
928665 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=928665]	Fix	When going back to overrides, directory is gone
923342 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=923342]	Fix	shutdown.jsp unable to compile
913876 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=913876]	Fix	ARCWriterPool timeouts -- legitimate?
896766 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896766]	Fix	If one URI connect-fails, hold queue, too
908719 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=908719]	Fix	Fetching simple URLs fails with S_CONNECT_FAILED (-2) error
809567 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=809567]	Fix	seeds held back/poor breadth first?
831480 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=831480]	Fix	Parsing links found between escaped quotes in JavaScript
895303 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=895303]	Fix	Does not extract applet URI correctky
791481 [http://sourceforge.net/tracker/index.php?func=detail&group_id=7	Fix	links to likely-embed types should be treated as embeds

ID	Type	Summary
3833&atid=539099&aid=791481]		
900826 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=900826]	Fix	Frontier.next() forceFetches will cause assertion error
877873 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=877873]	Fix	Flash link extractor causes OutOfMemory exceptions.
899976 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=899976]	Fix	Should be possible to resume from
896878 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896878]	Fix	Heritrix ignores charset
910210 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=910210]	Fix	Max # of arcs not being respected.
904723 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=904723]	Fix	New profile should ensure unique name
902940 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=902940]	Fix	When changing scope common scope settings are lost
903910 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=903910]	Fix	ssl doesn't work
903084 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=903084]	Fix	Allow that people use tcsh/csh not just bash
896788 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896788]	Fix	https SSLHandshakeException: unknown certificate

ID	Type	Summary
901397 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=901397]	Fix	Cannot override settings that isn't set in globals
892253 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=892253]	Fix	'Waiting for pause' even after all threads done
896800 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896800]	Fix	filter 'invert', filter names need work
896835 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896835]	Fix	max-link-hops (etc.) ignored unless
872069 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=872069]	Fix	order.xml absolute paths
892105 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=892105]	Fix	Cannot set TransclusionFilter attributes
874057 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=874057]	Fix	Link puts garbage into arc file: http://www.msn.com/robots.t

21. Release 0.6.0 - 2004-03-25

Release made in advance of radical frontier changes. Added bandwidth throttle, operator 'diary', settable robots expiration, crawler cookie pre-population, and changing of certain options mid-crawl. Many UI improvements including UI display of critical exceptions, UI desccription of job-order options, and improved reporting. Optimizations. Updated httpclient lib to 2.0 release and jmx libs to 1.2.1. Lots of bug fixes.

21.1. Changes

Table 14. Changes

ID	Type	Summary
861861 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=861861]	Add	861861 Redirects(/refreshes) from seeds should == new seeds
899223 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=899223]	Add	899223 Special seed-success report should be offered
891986 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=891986]	Add	891986 Bandwidth throttle function, setting.
877275 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=877275]	Add	877275 integrated operator 'diary' needed
891983 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=891983]	Add	891983 IP, Robots expirations should be settable
910152 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=910152]	Add	910152 Recovery of old jobs on WUI (re)start
781171 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=781171]	Add	781171 parsing css
912986 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=912986]	Add	912986 log views should give an idea of file size (where possible)
912989 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=912989]	Add	912989 Alerts should have 'select all' button...
856593 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=856593]	Add	856593 [load][save][turn on/off] cookies
912201 [http://sourceforge.net/tracker/ind	Add	912201 Add levels to alerts

ID	Type	Summary
ex.php?func=detail&group_id=73833&atid=539099&aid=912201]		
896665 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896665]	Add	896665 Split processor chains.
896754 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896754]	Add	896754 Show total of disregards
903095 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=903095]	Add	903095 Show increments of megabytes in ui
896794 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896794]	Add	896794 serious errors (eg outofmemory) should show up in UI
900520 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=900520]	Add	900520 Short description of ComplexTypes in user interface.
899982 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=899982]	Add	899982 Should be possible to alter filters while crawling.
896672 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896672]	Add	896672 Display progress (doc/sec) with more precision
896677 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896677]	Add	896677 Highlight the success or failures of each seed
896760 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896760]	Add	896760 Prominent notification when seeds have problems
896801 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896801	Add	896801 java regexps (in log view) need help text

ID	Type	Summary
]		
896778 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896778]	Add	896778 Log viewing enhancements:
896795 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896795]	Add	896795 frontier, thread report improvements
876516 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=876516]	Add	876516 default launch should nohup, save stdout/stderr
896763 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896763]	Fix	127.0.0.1 in job report
896767 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896767]	Fix	Frontier retry-delay should include units (eg -seconds)
898994 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=898994]	Fix	Revisiting admin URIs if not logged in should prompt login
899019 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=899019]	Fix	Deadlock in Andy's 2nd Crawl
767225 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=767225]	Fix	Better bad-config handling
815357 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=815357]	Fix	mysterious pause facing network (DNS) problem
896747 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896747]	Fix	ExtractorJS's report overstates it's discovered URIs
896667	Fix	Web UI does not display cor-

ID	Type	Summary
[http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896667]		rectly in IE
896780 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896780]	Fix	console clarity/safety
896655 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=896655]	Fix	Does not respect per settings added after crawl was started.
856555 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=856555]	Fix	'empty' records in compressed arc files

22. Release 0.4.1 - 2004-06-04

Memory retention fix.

22.1. Changes

Table 15. Changes

ID	Type	Summary
-	Fix	895955 [http://sourceforge.net/tracker/index.php?func=detail&aid=895955&group_id=73833&atid=539099]URIRegExpFilter retains memory

23. Release 0.4.0 - 2004-02-10

Release made for heritrix workshop, San Francisco, 02/2004. New MBEAN-based configuration, extensive UI revamp, first unit tests and integration selftest framework added, pooling of ARCWriters, new cmd-line start scripts, httpclient lib update (2.0RC3) and bugfixes.

23.1. Changes

Table 16. Feature

ID	Type	Summary
-	Add	New MBEAN-based configuration system. Reads and writes XML to validate against heritrix_settings.xsd.
-	Add	UI extensively revamped. Exploits new configuration system.
-	Add	60-odd unit tests added.
-	Add	Integration selftest framework.
-	Add	Added pooling of ARCWriters.
-	Add	Start script backgrounds heritrix and redirects stdout/stderr to
-	Add	876516 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=876516] Default launch should nohup, save stdout/stderr
-	Add	Web UI accesses are loggged to heritrix_out.log also.
-	Add	Updated httpclient to version 2.0RC3.
-	Add	763517 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=763517] IAGZIPOutputStream NPE under IBM JVM
-	Add	809018 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=809018] Cleaner versioned testing build needed
-	Add	872729 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=872729] Cmd-line options for setting web ui username/password
-	Add	863317 [http://sourceforge.net/tracker/index.php?func=detail&group_id=73833&atid=539099&aid=863317] Universal single-pass extractor

24. Release 0.2.0 - 2004-01-05

First 'official' release.

25. Release 0.1.0 - 2003-12-31

Initial Mavenized development version number (CVS/internal only). Added everything to new project layout.