Daemon Release Notes 1.54 – Current

Daemon 1.74.3

- **Bug Fixes**
  - Upgraded third-party libraries to address security vulnerabilities reported against them. Updated versions include Apache PDFBox 1.8.16 (CVE-2018-11797), Apache Commons Compress 1.18 (CVE-2018-11771) and FasterXML Jackson 2.9.7 (CVE-2018-7489).
  - Some of the ways ServeContent can be invoked failed in some cases on AUs having multiple crawl-start URLs, when some of the start URLs do not exist.

Daemon 1.74.2

- **Features**
  - The new metadata type "`File`" supports indexing of arbitrary publication types. Support is in place for both publication level items (`MetadataField.PUBLICATION_TYPE_FILE`) and article level items (`MetadataField.ARTICLE_TYPE_FILE`). Article level file items will be assumed to have a publication level file parent even if not explicitly defined. Item metadata beyond the standard access URL, publisher, and provider may be stored as arbitrary key-value pairs in a `MetadataField.FIELD_MD_MAP`.
  - Content Configuration web service now adds AUs from their TDB definition rather than by AUID, matching the way other subsystems add AUs: Including non-definitional parameters, and choosing the least full repository.
  - Deep crawl status information (`lastDeepCrawl`, `lastDeepCrawlResult`, `lastCompletedDeepCrawl`, `lastCompletedDeepCrawlDepth`) is tracked and reported in the UI, and through the `getAuStatus()` and `queryAus()` Web services.
  - Debug Panel and AU Status now include a "Validate Files" action which runs the plugin's `ContentValidator` on all files in the AU, reporting any `ValidationFailures` thrown.
  - In lieu of a MIME-type content validator factory, plugins may specify an `au_url_mime_validation_map`. `ValidationFailures` will occur for URLs that match one of the patterns but whose `Content-Type` does not match the corresponding MIME-type. E.g.,

```
<entry>
  <string>au_url_mime_validation_map</string>
  <list>
```

```
      <string>/doi/pdf(plus)?/, application/pdf</string>
      <string>/doi/(abs|full)/, text/html</string>
    </list>
  </entry>
```

- `ContentValidators` may throw `ContentValidationException.LogOnly` to record a warning message without causing validation failure.
- The “Files” list from AU Status now includes a `PollWeight` column.

- **Bug Fixes**
  - `SubscriptionManager` omitted non-definitional parameters when adding subscribed AUs.
  - The Link Rewriter rewrote in-page links (“`#ref`“), breaking them.
  - Metadata item type inference reversed `BOOKCHAPTER` and `BOOKVOLUME` in some circumstances.
  - In `queryAus()` web service, selecting `newContentCrawlUrls` field caused a fatal error.
  - The `LastMetadataIndex` field in `getAuStatus()` and `queryAus()` web services was not accessible using `daemonstatusservice.py`.
  - Fixed unsafe database resource closings and incorrect comparisons in metadata-handling code.
  - Fixed active task removal when metadata indexing for an AU is disabled.

Daemon 1.73.4

- **Features**
  - Allow the content configuration Web Service to use the same storage volume selection logic as the UI when adding AUs.
- **Bug Fixes**
  - Bug fixes in ServeContent link rewriting and OpenURL resolver.
  - Properly trigger configuration of AUs after synchronizing whole title subscriptions.

Daemon 1.73.3

- **Features**
  - The ViewContent screen now offers an option in the upper pane to run a link extractor on the content displayed in the lower pane.
- **Bug Fixes**
  - Fixed a bug in the title subscription management screen’s tabbed interface, which under some circumstances could cause the loss of title subscription data previously entered in other tabs.

Daemon 1.72.3

- **Features**
  - Plugins may supply a URL pattern to MIME map, with `au_url_mime_type`, which will be used to infer the MIME type of files that were collected without a `Content-Type` header. Also used for archive members whose extension isn't in the standard MIME map.
  - `ContentValidators` may now prevent (rejected) files from being stored, without causing the crawl to fail, by throwing a `ContentValidation` exception which the plugin maps to `CacheException.NoStoreWarningOnly` (or any other `CacheException` with `CacheException.ATTRIBUTE_NO_STORE` set).
  - In the case of redirects, a `ContentValidator` previously was invoked once for each URL in the redirect chain. Now it's invoked only once, on a `CachedUrl` representing the original URL, with the remaining URLs in the redirect chain stored in the properties, as the value of `x-lockss-validator-redirect-urls`.
  - AU poll and crawl priority can now be controlled based on arbitrary TDB attributes. Set `org.lockss.poll.pollWeightAuMap` or `org.lockss.crawler.crawlPriorityAuMap` to a list of `XPath,priority` pairs. See `AuXpathMatcher` and `AuXpathFloatMap`.
  - Per-peer lists of agreeing, disagreeing, etc URLs are available in the UI poll results, and the `queryPolls()` call in `DaemonStatusService`. This is currently somewhat memory intensive and must be enabled by setting `org.lockss.poller.v3.recordPeerUrlLists` to true.
  - Per-peer agreement values in poll results have always reflected the pre-repair state, even after repairs have been made. If the per-peer URL lists are enabled (`org.lockss.poller.v3.recordPeerUrlLists`), the per-peer agreement values are now updated to reflect the effects of any repairs.
  - Credentials (username/password) can be set for arbitrary collections of AUs, to facilitate collecting from password-protected sites such as Archive-It. If an AU's credential string in the TDB starts with "@", the remainder is interpreted as the name of a global configuration parameter, whose value will be used as the credential string.
  - If `org.lockss.proxy.ignoreIfModifiedWhenContentLengthWrong` is true (default false), when serving files whose stored `Content-Length` header differs from the actual size of the file, ignore any `If-Modified-Since` header in the request, so that the file gets served again. Workaround to recover from proxy bug below.
- **Bug Fixes**
  - The proxy could serve truncated files in situations where an incorrect `Content-Length` header was received from the origin server when the file was collected. The entire file is now served, with the original, incorrect file length as the value of the `X-Lockss-Orig-Content-Length` header.
  - `CrawlSeed`-supplied start URLs that don't match the crawl rules would still be fetched, but not parsed. They are now recorded as errors.
  - Fixed `ViewContent` malformed links that opened new tabs/windows spuriously.

Daemon 1.71.2

- **Features**
  - The daemon is no longer dependent on the configuration server in order to start. After the first startup (which does require the configuration server), the configuration is stored locally and will be used on subsequent startups if the config server is unavailable and the local copy is intact and not too old. This can be disabled, or a maximum acceptable age established, using hostconfig.
  - The crawler now sends Referer: headers (unless org.lockss.crawler.sendReferrer is set false). The referrer is remembered and will be sent if the file is subsequently repaired from the publisher (unless org.lockss.crawler.repair.sendReferrer is set false).
  - If a per-AU proxy spec (crawl_proxy param in tdb) is invalid, previously the system silently fell back to either the global crawl proxy setting, if any, or no proxy. Now the crawl will abort with an error. Set org.lockss.crawler.abortOnInvalidProxy to false to restore the old behavior.
  - Crawl Status page and result of queryCrawls web service include refetch depth and priority.
  - Crawl Status page and result of queryCrawls web service include all pending high priority crawls. Previously only the next several scheduled to run were included.
  - MetadataMonitor now displays a list of (former) AUs that have metadata recorded in the database but do not exist in the daemon, and allows the obsolete metadata to be deleted from the database.
  - Result of getAuStatus and gueryAus now includes lastMetadataIndex time.
- **Bug Fixes**
  - Crawl permission was erroneously found in some circumstances involving css on a permission page.
  - The referring page for permission probe pages was not displayed in the crawl status.
  - Crawls aborted due to certain unusual Java Errors were marked as Successful.
  - The NewFileVersion alert was improperly signalled when the crawler collected (but did not store) a file that was identical to the existing version.
  - Pending high-priority crawl requests (from DebugPanel or AuControlService) were lost when a new version of the AU's plugin was loaded.
  - Some metadata DB updates relating to item provider did not work with MySQL.
  - Publications with more than 2 ISSNs were not always correctly reported by MetadataMonitor.
  - Metadata Indexing Status page overreported the "Total Articles in Index" count after incrementally re-indexing existing AUs.
  - Exception mappings in plugin_cache_result_list rule did not take effect if the Exception class was defined in the plugin jar.
  - Due to an apparent change in behavior of some Java I/O classes, URLs stored in the LOCKSS repository have been limited to ~2K characters for some time. The limit has been restored to ~4K characters.
  - Upgrading the daemon rpm would overwrite local changes made to the logrotate configuration. Now the local changes are left in place; the new file is stored in /etc/logrotate.d /lockss.rpmnew .

Daemon 1.70.4

- **Bug Fixes**
  - This release addresses an incompatibility with OpenJDK 1.7.0_111. It is feature-equivalent to 1.70.3.

Daemon 1.70.3

- **Bug Fixes**
  - ServeContent sometimes served a file other than the one requested.

Daemon 1.70.2

- **Features**
  - A new validation framework allows plugins to validate content files upon collection and attempt to refetch (or not) if they don't pass. This is useful for sites that sometimes serve truncated or otherwise corrupted files, or transient HTML error pages in place of the real content, etc. See examples in `SamplePlugin.xml` and the LOCKSS javadoc.
  - The "Publications with multiple ISBNs/ISSNs" tool in `MetadataMonitor` now supports deleting selected ISSNs from the database.
  - The crawler now follows links contained in the content of data: URIs.
  - The proxy in use for each crawl, if any, is displayed in the crawler status.
  - Several additional composite tags added to default HTML parser used for filtering: `<audio>`, `<button>`, `<canvas>`, `<figcaption>`, `<figure>`, `<main>`, `<mark>` and `<video>`.
  - Peer Agreement tables and corresponding web service now include weighted agreement results.
- **Bug Fixes**
  - Fetching large files that took more than two hours to collect could cause daemon restarts.
  - ServeContent returned 404 instead of serving preserved content in some post-cancellation situations where the publisher site has been redesigned.
  - ServeContent returned 500 in some cases where the publisher served compressed content despite the absence of an Accept-Encoding request header allowing it to do so.
  - Java 7 and some other clients could get a `SSLHandshakeException: CertificateException` talking to the daemon's SSL proxy, if the proxy was configured to use a self-generated self-signed certificate. If this is happening, it's necessary to delete the old certificate and restart the daemon in order to force a new certificate to be created.
  - The SSL proxy could get into a bad state where all requests resulted in a 500 error with `IllegalStateException`.
  - During startup, the proxy erroneously returned 412 (Precondition Failed) instead of 503 (Service Unavailable) for requests that include an AUID, if the URL was not found.

- URL lookups during startup could cause erroneous “not found” results for those URLs for some time after startup.
- The metadata database records a `lastIndexedTime` for AUs even if indexing finds no articles, so they are included in query results.
- The RIS file metadata extractor didn’t properly process continuation lines.
- Value-less HTML attributes (expected to have values) could cause a `NullPointerException` in the HTML link rewriter.
- Unicode non-breaking space (U+00A0) in permission statement is now treated as whitespace.
- The list of start URLs in the crawl status was sometimes wrong for plugins that supply a `CrawlSeed`.
- Substance URLs and Files lists from AU status page didn’t always accurately reflect the substance status recorded for AU due to inconsistent handling of redirects. Substance lists now properly observe `org.lockss.substanceChecker.detectNoSubstanceRedirectUrl`
- Substance Files list contained duplicate items.
- Peer Agreement tables didn’t take reputation transfers into account.
- Some publishers could be missing from `DisplayContentStatus`.

Daemon 1.69.4

- **Features**
  - New AuControlService web service provides all the AU-related functionality of DebugPanel.
  - Added `tdbProvider` to queryAus web service queries and result.
  - Added `journalTitle` to queryAus web service result.
  - Many new items in MetadataMonitor servlet, and corresponding new methods in MetadataMonitorService service allow querying the metadata database for some common inconsistencies in extracted metadata.
  - New method `getAuUrls()` in DaemonStatusService web service enumerates URL subtrees in an AU.
  - The set of metadata fields required to be emitted by metadata extractors may be set as the value of `org.lockss.metadataManager.mandatoryFields`.
  - Subscription Manager supports subscribing to all current and future AUs, or to all current and future AUs belonging to specific publishers. Each publisher on the Title Subscription Management pages now has an Overall Subscription checkbox, and if `org.lockss.subscription.totalSubscriptionEnabled=true`, a new Total Subscription option will appear at the top.
  - The daemon’s link extractors, rewriters, metadata extractors, etc. now infer the character set used to encode HTML and other documents, in much the same way browsers do. Formerly the alleged charset (from the `Content-Type` header or the document itself) was always believed, but in many cases it’s not correct.

- The weight assigned to agreeing/disagreeing URLs in poll tallies may be controlled by plugins with `au_url_poll_result_weight`, a list of <pattern>,<float>. This does not affect the repair process, but may be useful to give a better indication of the preservation state and inter-peer agreement state of sites where auxiliary files such as CSS and JavaScript are rapidly versioned.
- Plugins may specify `org.lockss.util.None` for a `link_extractor_factory`, `crawl_filter_factory`, `hash_filter_factory` or `link_rewriter_factory` to remove the default factory for that MIME type.
- `org.lockss.crawler.globallyPermittedHosts` and `org.lockss.crawler.allowedPluginPermittedHosts` can be lists of patterns, or a single pattern as before.
- AU config params of type `Range` can now be a singleton: “N” is equivalent to “N-N”.
- New AU config param functors for use by plugins in `printf` arglists: `range_min()`, `range_max()`, `num_range_min()`, `num_range_max()` can be used with params of type `Range` and `NumRange`.
- Reduced volume of crawl end alert message by allowing CrawlFinished and CrawlFailed alerts to be grouped a single alert email independent of AU. Set `org.lockss.alert.specialGroups` to a list of <AlertName>,<GroupName> (default `[CrawlFinished,CrawlEnd, CrawlFailed,CrawlEnd]`). If <GroupName> starts with “`au:`“, alerts must have matching AU to be grouped; if starts with “`plugin:`“, alerts must have matching plugin to be grouped.
- `plugin_parent_version` is now checked when loading a child plugin. The action taken on mismatch is controlled by `org.lockss.plugin.parentVersionMismatchAction`, one of Ignore, Warning (default) or Error (throws `PluginException.ParentVersionMismatch`).
- The last config reload time is displayed in DaemonStatus / Configuration.
- If an AU has been assigned a crawl priority or its crawls have been disabled (via `crawlPriorityAuidMap`), that fact will be displayed on the AU status page.
- If `org.lockss.baseuc.proxyByAuid` is true, proxied crawls will include an `X-Lockss-Auid:` header in the request, specifying the AU being crawled. When the proxy receives this header it will serve files only from the specified AU.

- **Bug Fixes**
  - The automatic decompression upon receipt of (unbidden) gzipped and deflated content introduced in 1.67 has been removed. The previous behavior violated one of our preservation principles (preserve exactly what is collected) and created several other problems. Compressed files are now stored that way, and uncompressed when appropriate depending on use. This should be transparent to users, except those who access raw files in the repository.
  - ServeContent did not work properly with URLs containing unicode characters.
  - HTTP headers containing non-ascii characters were incorrectly encoded. The encoding now defaults to ISO-8859-1 (and can be changed with `org.lockss.urlconn.httpHeaderCharset`

but there should be no reason to do so).
- By default, the crawler retries fetches when the length of the received content disagrees with the `Content-Length` header. May be overridden in plugins by mapping `ContentValidationException.WrongLength` to `CacheException.WarningOnly` (to cause a warning) or to `CacheSuccess` (to completely ignore inconsistency).
- Ensure no file is stored after a validation failure in `DefaultUrlCacher`.
- Plugins may use `au_additional_url_stems` to inform the daemon that they collect URLs with stems that don't match any of their config params. This is necessary in order for those files to be served.
- Implicit permission granted by `org.lockss.crawler.globallyPermittedHosts` and plugin-specified permitted CDN hosts failed to apply to start pages.
- Permission pages redirected to excluded URL, and failure to fetch permission page are reported correctly in the crawl status.
- DebugPanel / Reindex schedules the AU to be (re)indexed before those added to the queue automatically (due to new crawl or changed metadata extractor).
- Fixed bugs in the display of the preserved volumes in the Keepers report and added an option to enumerate all preserved volumes instead of coalescing into ranges.
- Regexps in AU definition are now displayed correctly in DaemonStatus / AuConfiguration.
- Not all publishers were displayed in the DisplayContentStatus view.
- Footnotes were sometimes missing from SubscriptionManagement UI pages.
- Failed crawls could erroneously be recorded as successful after unusual JVM errors.
- Fixed filedescriptor leaks caused by following redirects in SubstanceChecker.
- COUNTER aggregated statistics are included in config backup file.
- Proxy sent incorrect `X-Forwarded-For for https:` requests.
- Proxy sometimes returned truncated files for `https:` requests.
- Ensure the accounts directory is not world-readable.
- `hostconfig`-created admin password is encrypted with SHA-256.
- `hostconfig` sets permissions of toplevel cache directories to 750.
- Log rotation for daemon log didn't take effect with recent versions of logrotate.

Daemon 1.68.4

- **Features**
  - The ContentService web service has new methods:
    - `getVersions(String url, String auid)`: Returns a collection of objects containing the version number, the file size and the collection date.
    - `fetchVersionedFile(String url, String auid, Integer version)` is similar to the existing `fetchFile(String url, String auid)` but returns the specified version instead of the latest one.
  - The result of `ContentService.fetchFile()` includes the `CachedUrl` properties (response

headers, etc.)
    - The lists of URLs (including size and number of versions), substance URLs and article URLs are now available from `DaemonStatusService.queryAus()`
    - A link extractor for XML is now included, and the crawler follows links to style sheets in XML files.
    - If `org.lockss.baseuc.socketKeepAlive` is `true` (default `false`), `SO_KEEPALIVE` will be enabled for crawler TCP connections. This should eliminate daemon exits resulting from hung connections due to server or network outages. The default will be changed to `true` in an upcoming release.
    - Records in the metadata database with `Unknown` provider are updated on daemon startup if/when the provider is added to the TDB.
    - Experimental support for using Shibboleth to authorize UI access is available.
    - A new `Alert`, `NewFileVersion`, is raised when a different version of file that already exists is collected.
    - A new `Alert`, `FileVerification`, is raised when a file is collected whose size disagrees with the `Content-Length` response header.
    - If `org.lockss.thread.exitDaemonOnOome` is `true` (default `false`), threads that exit due to an otherwise unhandled `OutOfMemoryError` will cause a daemon exit. If `triggerWDogOnExit(true)` has been called then `threadExited()` will be called instead so the thread can handle the situation.
- **Bug Fixes**
    - The major causes of temp files accumulating over time have been fixed.
    - Files collected from CDN hosts or globally permitted hosts (e.g., those on hosts for which the plugin doesn't explicitly list a permission page) formerly wouldn't be found when searched for by URL.
    - Requesting metadata indexing from `DebugPanel` didn't always result in a full reindex.
    - URLs and keywords extracted during a reindexing operation now replace existing URLs and keywords in the database. Formerly they were added to existing data.
    - Metadata extraction made an invalid assumption that an AU contains content from a single publisher. This is not necessarily true, especially of file-transfer AUs. Each metadata item is now associated with its own publisher.
    - AUs with no corresponding TDB entry caused errors when upgrading the metadata database to version 22.
    - Fetch errors in repair crawls were incorrectly reported as "Unexpected error: This crawler has no failed urls set".
    - Crawls aborted because a new version of a plugin was loaded had a blank status.
    - Redirects to a login page did not immediately abort the crawl.
    - If a start URL was redirected to a URL excluded by the crawl rules, the crawl failed with no explanation.
    - Multiple response header fields with the same name were not handled properly by the crawler

– only the last value was stored. The multiple values are now concatenated into a comma-separated list (see RFC 2616). The parameter `org.lockss.urlconn.singleValuedHeaders` may be set to a list of header names that will be treated the old way.
- URLs with query arguments containing slashes could be skipped by iterators in some situations.
- UI authentication failures now incur a delay.
- `JsoupHtmlLinkExtractor` emitted spurious links in some cases when plugins registered handlers for additional attributes.
- The status of asynchronous HashCUS hashes that were aborted due to `OutOfMemoryError` appeared to be `Running` forever.
- The AU agreement value is now always included in `AuWsResult`.

Daemon 1.67.6

Released only to CLOCKSS.

- **Features**
  - DebugPanel has a Check Substance button that can be used to rescan for substance AUs that improperly got marked as having no substance, and which can longer be crawled.
- **Bug fixes**
  - The crawler no longer marks AUs as having no substance just because the manifest has been changed to no longer link to previously-collected substance. A new alert (`NoSubstantialContentLinked`) is raised in this case and the AU is properly marked as having substance.
  - Fixed parsing problems in OAI-PMH responses.
  - Fixed regression from PDF framework NPE fix in 1.67.4.

Daemon 1.67.5

Please also see the 1.67.4 release notes below.

- **Features**
  - Several improvements to the CrawlSeed framework:
    - Added the `BaseCrawlSeed(CrawlerFacade)` constructor
    - Made `BaseCrawlSeed.getStartUrls()` and `getPermissionUrls()` final, added overrideable `doGetStartUrls()`, `getPermissionUrls()` and `initialize()` methods.
    - Added `makeUrlFetcher()` to `CrawlerFacade`.
- **Bug fixes**
  - Lists of URLs, Articles, Metadata, etc. from the links on the AU status page (the output of ListObjects) are now UTF-8 encoded.

Daemon 1.67.4

Released only to CLOCKSS.

- **Features**
  - The pattern for recognizing RIS tag lines in RisFilterReader is now customizable with `setTagPattern(Pattern)`.
  - Creative Commons licenses with `https:` URLs are now recognized.
  - Tdb entries may specify a provider, distinct from the publisher.
- **Bug fixes**
  - The `CrawlerFacade` is now passed to `CrawlSeedFactory.makeCrawlSeed()`
  - Improved Crawler error consistency.
  - Fixed an intermittent problem where ServeContent, given an archive member name or DOI, served the entire archive instead of just the member.
  - Full metadata reindexing (due to plugin change or requested from DebugPanel) did not erase previous info for AU nor descend into previously indexed archive files.
  - Removed an invalid assumption in the metadata extraction framework that all content in a single AU belongs to the same publisher.
  - Added `<center>` as a composite tag name in HtmlParser
  - Issue range AU config parameters with leading zeroes were interpreted as octal (since 1.67.3).
  - The daemon didn't always restart after the first exit following an upgrade.
  - `AU.getStartUrls()` caused NPE in plugins with a crawl seed.
  - Proxy requests with `If-Modified-Since: Thu, 01 Jan 1970 00:00:00 GMT` resulted in `304 Not Modified` for cached files with no recorded Last-Modified.
  - Fixed NPE in string decoding and font handling code in PDF framework.
  - Fixed NPE displaying hash queue when time-recalc hashes are present.

Daemon 1.67.3

- **Features**
  - **Crawler**
    - The crawler and parts of the plugin framework have been refactored to allow more flexibility in interacting with a site to determine what to collect and how to process received content. Plugins may specify or supply a `CrawlSeed` to perform, e.g., metadata queries to obtain either a complete list of URLs to collect, or starting points. The old OAI-PMH crawler has been incorporated into this framework, and a DSpace plugin will be available soon.Content may be processed when fetched by specifying a `UrlConsumer`. The Exploder framework, which unpacks archives on collection, has been recast as a `UrlConsumer`.There are two changes affecting probe permission pages (the URL used to test for subscription access when the manifest page is not access-controlled): the probe URL is now subject to the crawl rules – the probe will fail if

the URL doesn't match. And the probe page is now stored in the AU by default. To suppress this (e.g., for sites where the probe page doesn't belong in the AU), plugins may set `plugin_store_probe_permission` to `false`.

- PLN admins and plugin writers may now assert that permission has been granted to collect content from hosts without an explicit permission statement or Creative Commons license. This intended for situations in which it's impractical to get a LOCKSS permission statement or CC license added to a site – it cannot be used to circumvent subscription control.To support collection from distribution sites for standard css, js, etc. libraries, `org.lockss.crawler.globallyPermittedHosts` may be set to a regexp – matching hostnames will be permitted. E.g., many sites load the jquery library from `ajax.googleapis.com` – to permit the LOCKSS crawler to collect jquery and other similar libraries, `globallyPermittedHosts` may be set could be set to `ajax\.googleapis\.com`.To support content distribution networks, plugins may set `au_permitted_host_pattern` to a (list of) printf templates for regexps. Matching hostnames will be permitted, but only if they also match the regexp `org.lockss.crawler.allowedPluginPermittedHosts`; this gives the PLN admin control over the hosts for which plugin writers may assert permission.
- The `Accept:` header sent by the crawler in HTTP requests may be set with `org.lockss.urlconn.acceptHeader`. The default is `Accept: text/html, image/gif, image/jpeg; q=.2, */*; q=.2`. This is a change from the previous value: `Accept: text/html, image/gif, image/jpeg, *; q=.2, */*; q=.2` which was found to trigger errors on some servers.
- Plugins can direct the crawler to add arbitrary headers to the HTTP requests it sends by setting `au_http_request_header` to a list of `:` strings.
- PLN admins can globally direct the crawler to add headers to the HTTP requests it sends by setting crawler by setting `org.lockss.crawler.httpRequestHeaders` to a list of `:` strings. Any headers specified by plugin with `au_http_request_header` take precedence over this.
- The crawler now handles gzipped & deflated response bodies even though it doesn't send an `Accept-Encoding:` header indicating its willingness to accept them. Some servers erroneously send gzipped responses even when not requested.
- Plugins can direct the crawler to perform a depth-first traversal of the links it finds in pages instead of the default breadth-first, by setting `plugin_crawl_url_comparator_factory` to `org.lockss.plugin.DepthFirstCrawlUrlComparatorFactory`.
- If `org.lockss.baseuc.stopWatchdogDuringPause` is `true`, the crawler's thread watchdog will be disabled during pauses imposed by the fetch rate limiter. This is necessary if one or more plugins use extremely long fetch delays as a crawl windowing technique.

- **Poller**

- Since release 1.62 the daemon has included support for Proof of Possession and Local polls in addition to the traditional Proof of Retrievability polls. Proof of Retrievability polls hash the entire content of an AU. Proof of Possession polls hash a random sample of an AU's content and are a less expensive way for peers to prove to other peers that they share the same content for the AU. Local polls compare the computed and stored hashes of the content of an AU as a low-cost hint as to whether the AU needs a Proof of Retrievability poll to repair possible damage.We have tested these new poll types and plan shortly to enable them in the GLN and CLOCKSS networks. To support this 1.67 includes a policy that determines when peers will call these new types of poll.
- To help deal with sites that frequently version css and other ancillary files, plugins may set `au_repair_from_peer_if_missing_url_pattern` to a (list of) regexps. During polls, URLs that match will be fetched from a peer if not present on the poller, even if the vote is too close for the normal repair mechanism to conclude that the file should exist.

- **Other**
  - A new web service, `HasherService`, allows clients to request the hash of one or more preserved files, akin to the HashCUS servlet. See http://:8081/ws/HasherService?wsdl for details.
  - A new web service, `ContentService`, allows clients to fetch the contents of a preserved file. It has a single method, `fetchFile()`, which takes a URL and AUID.
  - The web services queries pertaining to crawls, polls and votes now include the AUID and/or crawl key in results, and allow it in queries, where meaningful.
  - The Java runtime version is available in the UI in Platform Configuration, and in the DaemonStatus web service as `javaVersion` in the result of `getPlatformConfiguration()`.
  - Leading and trailing whitespace is removed from metadata text fields. Existing databases are updated to this state when this release first runs.
  - AUs whose metadata has not yet been indexed will be indexed before reindexing other AUs.
  - The concept of a "provider" has been added to the metadata database and subscription records, to handle the situation where the same title is available from multiple sources (e.g., the publisher and an aggregator).
  - Publications may now have multiple proprietary identifiers, supplied by different providers. These appear in COUNTER reports.
  - printf args in plugins can now invoke functions, either built-in or defined by the plugin. E.g., plugins can handle sites that mix hostnames with and without "www." by specifying additional crawl rules and permission pages using `add_www(base_url)` or `del_www(base_url)`.The following functions are built-in:
    - `url_host` – returns the host part of a URL
    - `url_path` – returns the path part of a URL
    - `add_www` – prepends "www." to the host part of a URL if not already present

    - `del_www` – removes leading “www.” from the host part of a URL
    - `to_https` – replaces “http:” scheme of a URL with “https:”
    - `to_http` – replaces “https:” scheme of a URL with “http:”
    - `url_encode` – URL-encodes its argument
    - `url_decode` – URL-decodes its argument

    Plugins may define additional functions by setting `au_param_functor` to the name of a class that implements `AuParamFunctor` or extends `BaseAuParamFunctor`.

  - `plugin_au_config_user_msg`, which causes a plugin-specifiable message to be displayed when a user adds AUs, can now be a printf template.
  - The AU status page now has a link to the AU’s browseable content (ServeContent).
  - The config backup file is now generated periodically at a predictable location on disk and may be fetched or backed up from there. The Backup link in Journal Configuration uses this file if present so doesn’t incur the delay to generate the file. By default the filename is `/cache0/gamma/backup/config_backup.zip` (where `cache0` is the first configured disk). The name may be changed by setting `org.lockss.backup.fileName`, the directory by setting `org.lockss.backup.dir`.
  - The hash files created by HashCUS have always been automatically deleted after they’re fetched. This may now be suppressed by setting `org.lockss.hashcus.autoDeleteHashFiles` to `false`.

- **Security Enhancements**
  - The SSL protocols SSLv3 and SSLv2Hello, recently discovered to be insecure, have been disabled by default.To reenable or disable a different list of protocols for HTTPS connections to the UI or content server, set `org.lockss.ui.disableSslServerProtocols` and/or `org.lockss.contentui.disableSslServerProtocols` to `null` (to re-enable all) or to a different list of protocols to disable.To re-enable or disable a different list of protocols for poll messages (LCAP-over-SSL), set `org.lockss.scomm.disableSslClientProtocols` and/or `org.lockss.scomm.disableSslServerProtocols` to `null` (to re-enable all) or to a different list of protocols to disable.
  - ImageIO native code has been disabled to close off a possible attack vector. The ImageIO library may be invoked by the PDF filtering framework – the equivalent Java library will now be used.

- **Bug fixes**
  - The type of the AuWsResult property `tdbYear` has been changed from integer to string as the value may be a range. Web services clients that call `DaemonStatusService.queryAus()` will need to be recompiled and may require some code changes.
  - HashCUS/List Requests failed if any AUs with pending or completed requests had been deleted.

- The space character was erroneously accepted as a legal separator between printf arguments in plugins. Comma is now required.
- ServeContent sometimes served the 404 – not found AU index page with a 200 response code, causing the index to be displayed in inappropriate places such as embedded images.
- When re-fetching URLs whose current preserved content is an empty file, an If-Modified-Since header was erroneously sent.
- `ContentDmLinkExtractor` was broken by a recent change to `DublinCoreLinkExtractor` to avoid repeatedly fetching dTD resources from the net. `ContentDmLinkExtractor` is no longer necessary – use `DublinCoreLinkExtractor` instead.
- Displaying the AU configuration of a plugin registry AU caused a `NullPointerException`.
- If no votes were present in a symmetric vote message the voter encountered an error and aborted.
- Daily crawl windows were one hour off during daylight savings time.
- The `RisFilterReader` ignored some RIS tags because it didn't know that the second character of RIS tags can be a digit.
- When trying to match incoming metadata to a publication already in the database, mismatches between existing and incoming ISBN/ISSN now cause a new publication to be created. Formerly, different publications with the same title might be stored together.
- The preserved count in List Titles was inaccurate for bulk content. The count in the metadata database is now used if present.
- The last content change timestamp on AUs is no longer updated if the only change is to the manifest page. This should prevent unnecessary metadata indexing, and is necessary for the poll policy logic.
- The PDF filtering framework got a `NullPointerException` on blank pages that don't have an internal PDStream.
- SubscriptionManager allowed subscriptions to be created to titles all of whose AUs were marked down.
- COUNTER Reports logic for determining whether the publisher had been contacted incorrectly included requests that attempted but failed to contact the publisher.
- Backup files sometimes could not be generated due to a mixup between print and electronic ISSNs.
- Database version updates that didn't complete on daemon upgrade are now retried until successful.

Daemon 1.66.3

- Features
  - Additional Web Services operations to:
    - Query Tdb entries
    - Add and delete AUs

        - Deactivate and reactivate AUs
    - Preserved files whose URL doesn't match the plugin's current crawl rules (because those rules have changed since the file was collected) are now excluded from normal processing by default. (Except that they're included in the node list on the AU status page, and ViewContent notes that they're hidden from normal processing.) Set `org.lockss.baseCachedUrl.includedOnly` and `org.lockss.cuIterator.includedOnly` to false to have them included in processing as before.
    - Empty files collected during a crawl are now reported in the crawl error list. If `org.lockss.crawler.refetchEmptyFiles` is true, empty files are refetched during recrawls independent of depth.
    - The level of trust required to collect files from HTTPS servers can be selected by setting `org.lockss.urlconn.serverTrustLevel` to one of:
        - Trusted: server certificate must be signed by a known CA.
        - SelfSigned: server certificate must be self-signed or signed by a known CA.
        - Untrusted: any server certificate will be accepted.

      The previous behavior was equivalent to SelfSigned. The default is Untrusted, as that's no less trustworthy than SelfSigned.

    - TCP keepalives can be enabled for UI connections by setting `org.lockss.ui.keepAlive` to true. This will allow long UI operations to complete without connections being closed prematurely. (On most systems it's also necessary to use `sysctl` (or modify `/etc/sysctl.conf`) to reduce `net.ipv4.tcp_keepalive_time` to less than 60 minutes.)
    - A new alert, `PluginNotLoaded`, is raised when a loadable plugin fails to load or start due to an error.
    - The `CrawlEnd` alert includes the proxy setting, if any.
- Bug fixes
    - OpenURL requests including both DOI and bibliographic parameters could fail even though content is present, if the content is covered by more than one plugin.
    - Deleted AUs are now removed from any active subscriptions, so they stay deleted.
    - Inactive AUs are excluded from metadata query results.
    - Spurious local hash disagreements have been eliminated. (Experimental feature.)
    - URLs that match the crawl rules but are excluded during a crawl because they redirect to a URL that doesn't match the crawl rules are now reported clearly in the crawl exclusion list.
    - RIS file filtering is more robust.
    - Parsing RDF (in CONTENTdm metadata, CreativeCommons licenses, etc.) no longer fetches standard DTD files from the network.
    - The status of polls aborted because hashing isn't completed by the deadline is reported as "Expired" instead of "Error".
    - Failure to load a plugin because its parent can't be found is now reported clearly.
    - Null values in AU configuration parameters could cause errors in Web Services queries.

Daemon 1.65.5

- Features
  - Initial implementation of a Web Services API. Documentation is available at http://plnwiki.lockss.org/wiki/index.php/Web_Services
  - The HTML parsers now recognize the HTML 5 encoding declaration (‘`<meta charset="XXX" />`‘).
  - CSS embedded in style attributes of HTML tags is processed for link extraction, and rewriting by ServeContent.
  - Links in “`onclick`” attributes are rewritten by ServeContent.
  - Tdb info is included in the AU Configuration page.
  - If `org.lockss.poll.v3.minReplicasForNoQuorumPeerRepair` is greater than 0 (default -1) voter-only URLs with too few votes for landslide, but which have agreeing plain hash from at least this many peers, will be repaired from one of those peers.
  - The Repair Candidates table in the UI by default reflects agreement recorded by any type of poll (Proof of Retrievability, Proof of Possession and symmetric versions thereof). With the query arg “`polltype=XXX`” (one of “`POR`“, “`POP`“, “`SYMMETRIC_POR`“, “`SYMMETRIC_POP`“) it displays agreement recorded only by that type of poll. A new Peer Agreements table displays the raw agreement data for each type of poll.
  - In proxy requests, the `X-Lockss-Local-Addr:` header instructs the proxy to bind its outgoing socket to the specific local address. Useful to manage publisher bandwidth limits on multi-homed boxes. The header takes effect only if `org.lockss.proxy.allowBindLocalAddresses` is set to either “`*`“, “`ANY`“, or a list that contains the requested address.
  - The subscription database is backed up and restored as part of the config backup.
  - More article metadata (volume, issue, start page) is exported to the external reporting database.
  - The list of “substance” URLs (those satisfying the plugin’s substance patterns or predicate) is available from the AU status page.
  - When testing in a development environment, it’s no longer necessary to list all plugins in `org.lockss.plugin.registry`. Instead set `org.lockss.plugin.registryJars` to a list of (unqualified) names of jars on the classpath (e.g., “`lockss-plugins.jar`“) and all plugins (whose name ends with “`Plugin`“) in the listed jars will be loaded. Plugins whose name doesn’t end with “`Plugin`” still need to be listed individually in `org.lockss.plugin.registry`.
- Bug fixes
  - In PLNs using LCAP SSL, transient network errors could cause polling communication to become degraded or stop over time, requiring daemon restart.
  - Memory leak in crawler caused daemon to grow over time.
  - Memory leak in PDF filtering framework caused daemon to grow over time.
  - NPE tallying symmetric vote when more than one version hashed.

- COUNTER reports can handle the (unusual) case where two journals have the same name and publisher.
- Poll agreement hints were incorrectly recorded as symmetric poll agreement hints.
- AU table incorrectly truncated that name of plugins that contain a dot.
- Substance is (re)checked during polls if the plugin's substance checker feature version has changed. Previously, down AUs would never be rechecked even if the patterns were updated.
- The poller wrongly sent a symmetric vote for AUs without substance.
- OpenURL requests resulting in 404 responded with an unsorted list of candidate AUs.
- The AU status table didn't display the substance state for AUs whose plugins supply a `SubstancePredicate` in place of patterns.
- `repairFromPublisherWhenTooClose` could cause collection attempts from publisher even if the AU was marked down.
- Enabling/disabling the `FetchTimeExporter` no longer requires daemon restart.
- Changing `org.lockss.plugin.registries.crawlProxy` no longer requires daemon restart.
- Proxy and local bind address settings could leak between different clients of connections to the same host.
- `org.lockss.crawler.crawlFromAddr` did not take effect.

Daemon 1.64.3

- Features
  - The proxy supports SSL connections and can serve files with https: URLs.
  - The proxy's 404 page is customizable. Set `org.lockss.proxy.errorTemplate` to the name of the file containing the template; see an example here.
  - ServeContent rewrites URLs in selected html meta tags of the form (commonly used for citations, etc.) Plugins should set `plugin_rewrite_html_meta_urls` to the list of names whose corresponding URLs should be rewritten.
  - URLs in meta tags will be rewritten as absolute if `org.lockss.htmlRewriter.metaTagRewritePrefix` is set to a URL prefix. This is a special feature for Google Scholar, which requires absolute links in citation metadata.
  - ListObjects output includes the number of items.
  - If `lockss.metadata.preferTdbPublisher` is true, the publisher stored with each article's metadata will be taken from the title db (if present) instead of the value obtained by the ArticleMetadataExtractor.
  - The crawler accepts CreativeCommons V4.0 licenses.
  - The set of valid CreativeCommons license types and versions is configurable.
  - The ExpertConfig textarea is resizable and initialized to a reasonable size.
- Bug fixes
  - URLs containing unicode characters can now be collected.
  - Password hashes are no longer included in the configuration email sent by hostconfig.

- Journal Configuration / Synchronize Subscriptions could fail if there were entries in the title db with no year, volume or issue.
- Triggering metadata reindexing from DebugPanel did not perform a complete reindex of the AU, potentially resulting in missing metadata.
- Metadata indexing performance has been improved, esp. for AUs where the metadata is extracted in archive file members.
- Metadata indexing of AUs with archive files didn't promptly close files, leading to "Too many open files" errors.
- Some parts of the audit proxy configuration didn't take effect if the main proxy was not configured.
- The syslog logger and alert action misinterpreted the configured facility.
- ServeContent returned 404 when attempting to access an archive member whose name contains spaces.
- Changing the substance pattern feature version in a plugin didn't always cause the substance state to be recomputed during votes.
- PDF filters could throw errors due to missing fonts.
- Access logs and alerts omitted client IP address.

Daemon 1.63.2

- Features
  - Added JsoupTagExtractor(Factory), a robust HTML and XML metadata extractor based on jsoup.
  - More complete auditing of user actions and configuration changes.
  - Hash status displays speed of individual hash requests.
- Bug fixes
  - The subscription manager adds in-range AUs only if they're still available from the publisher.
  - A malformed tdb entry could cause all tdb entries for that plugin to be ignored.
  - Missing title info or metadata no longer causes failed database updates.
  - Performance of metadata indexing with PostgreSQL is markedly improved.
  - The repository was unable to store files with long URLs containing unicode characters.
  - SimpleHtmlMetaTagMetadataExtractor missed some common cases due to whitespace variations.
  - DisplayContentStatus servlet: shift-click selection now works, plus speed and reliability improvements.
  - Title List (KBART) report: Improved reporting of contiguous holdings ranges using volume and year values, including for titles that have no volumes or non-numeric volume names.
  - The run_dev testing framework no longer loads plugins from lockss-plugins.jar, so properly reproduces plugin-loading behavior of production daemon.

Daemon 1.62.4

- Features
  - The proxy includes audit-related properties (checksum, repair info) in response headers if org.lockss.proxy.includeLockssAuditProperties is true (default false).
  - PostgreSQL is supported in addition to Derby (for Metadata, COUNTER and Subscription info).
  - The Subscription manager distributes added AUs across available disks in proportion to available space.
  - MetadataManager synthesizes publication names when missing from the generated metadata. Previously the incomplete records were not stored at all or caused errors.
  - SimpleHtmlMetaTagMetadataExtractor normalizes whitespace to a single space char.
  - FormFieldRestrictions can exclude entire forms (by id, name, action, submit value) from link extraction.
  - Many improvements to symmetric polling.
  - Initial implementation of “local” polls, which use local hashes to more quickly detect corruption and focus full polls where they are most needed.
  - Initial implementation of Proof of Possession (PoP) polls, which establish rights to receive repairs from peers with less work than full polls (now called Proof of Retrievability polls).
  - HashCUS includes hash times in background task list.
  - Integration with FITS (File Information Tool Set) is available, but currently requires a special build.
- Bug fixes
  - When serving cached content, the proxy now includes all the response headers that were originally received with the file.
  - DisplayContentStatus pages no longer display unformatted output before switching to the proper format.
  - Some URL order comparisons that should be consistent with URL traversal order weren’t, potentially leading to incorrect poll behavior.
  - Added database indices to minimize deadlocks between indexing and COUNTER reports aggregation operations.
  - Fixed an intermitten ConcurrentModificationException displaying CrawlStatus URL lists.
  - GoslingHtmlLinkExtractor allows the base URL to be set only once (matches behavior of browsers and other link extractors).
  - Plugin key au_crawl_depth has been renamed to au_refetch_depth. The old name will continue to be supported indefinitely. au_refetch_depth specifies which previously collected files should be refetched (checked for changes) on subsequent crawls.
  - Dashes are allowed in bibliographic years, volumes and issues.
  - Incremental improvements in daemon startup time with a large number of AUs.
  - Improvements in SubscriptionManager performance due to improvements in bibliograph

utilities.
- Additional memory savings from elimination of duplicate strings.
- Fixed a deadlock that caused the admin UI to become unresponsive.
- Fixed a race condition that caused stf tests to fail.

Daemon 1.61.6

- Bug fixes
  - The poller failed to fetch repairs from the publisher for some files, depending on the level of agreement between the voting peers.
  - In unusual circumstances, the poller could attempt to fetch repairs from a publisher that was marked “down”.
  - ListObjects (URL, file, DOI, metadata, etc. lists) output was sometimes truncated.
  - Inconsistent normalization of URLs containing backslashes could cause errors while enumerating the files in an AU.

Daemon 1.61.5

- Features
  - Symmetric polling: The poller now exchanges hashes bidirectionally with each voter, potentially doubling the rate at which nodes may prove possession and thus become willing repairers for each other.
  - HashCUS now allows client tools to enqueue hash requests and poll for completion, instead of blocking for the duration of the hash.
  - The Title List KBART report now shows values for the “coverage depth” field in all rows, currently taking the values “abstracts” or “fulltext”.
  - Exceptions to the set of config params disallowed by Expert Config may now defined by setting both org.lockss.config.expert.deny and org.lockss.config.expert.allow.
  - More database optimizations to better handle large amounts of metadata.
  - If remote access to the metadata database is enabled, it is password-protected. (Generally used only for development.)
  - Test coverage reports have been restored to the build.
- Experimental
  - Subscription Manager. Support has been added for automatically configuring new AUs for selected (“subscribed”) serial publications. It’s disabled by default in this release as there are some unusual numbering schemes present in the GLN that aren’t yet handle correctly. Feedback is welcome: set org.lockss.subscription.enabled=true and restart.
- Bug fixes
  - Some ebook titles were showing in the “journals only” report. The journals and books reports are now complementary.

- Publications with multiple names displayed the wrong totals in COUNTER reports.
- Some ISBNs and ISSNs that were part of the metadata of some AUs, but were ignored, will now stored in the database.
- Content requests from other LOCKSS crawlers are now filtered out of COUNTER statistics.
- During startup, the content server and proxy could return 404 for URLs that are preserved and would later return 200. They now return 503 in this situation, to indicate that the client should try again.
- A wider variety of unusual behaviors by plugin article iterators and metadata extractors is handled.

Daemon 1.60.3

- Features
  - Content accessed via form submission can now be collected. Simple cases (checkboxes, radio buttons, menus) can be handled automatically (by enumerating possible inputs); plugins may provide helpers to constrain the combinations, or supply values for text input fields.
  - Memento support has been added. The daemon implements TimeGate and TimeMap services, allowing historical versions of content to be viewed from Memento-aware browsers.
  - A Sitemap parsing & link extraction library has been added.
  - Plugins may specify URLs of files whose content (or existence) will be ignored in polls, by setting au_exclude_urls_from_polls_pattern to a regexp template or list thereof. Useful for pages that are expected to be served with different content each time they're fetched, and aren't amenable to hash filtering.
  - Polling priority can be influenced by setting org.lockss.poll.pollWeightAuidMap to a map of AUID regexp to priority. AUs are assigned the corresponding priority of the first regexp that their AUID matches.
  - The Metadata Indexing Status table has been enhanced to retain a list of failed indexing operations, and provide detail tables that show diagnostic information about the problem.
  - Plugin-supplied URL normalizers may now change the stem (host & port) only within the set of stems on which the AU is known to have content (based on start URL(s) and permission URL(s)). Arbitrary stem changes are once again illegal and org.lockss.UrlUtil.unrestrictedSiteNormalize has been removed.
  - Metadata indexing is much faster due to the addition of indices to several database tables.
- Bug fixes
  - HTML pages that declare their character encoding in a `<meta>` tag were sometimes rendered incorrectly by ServeContent.
  - Inconsistent URL normalization prevented some preserved pages from being found when viewing content.
  - ServeContent didn't add required Via header when forwarding requests to publisher.
  - The proxy erroneously served Last-Modified: Thu, 01 Jan 1970 00:00:00 GMT when no last-

modified date was known.
- A bug in the SFX DataLoader output that was introduced in 1.59 has been fixed. The correct ISSN field is now being generated.
- Counts of AUs available to add were incorrect on Add Titles pages if deactivated AUs were present.
- Permission checkers were failing to find permission in some cases due to bugs in the Boyer Moore string search implementation. Java regexp matching is now used instead.
- The crawl queue builder was using excessive CPU time in certain situations on boxes with tens of thousands of AUs.
- A race condition accessing UI session state occasionally led to looping and excessively high load avg.
- A serious performance problem accessing pages with opaque URLs, in situations where thousands of AUs contain content on a single host, has been mitigated.
- Some metadata extracted from HTML pages was improperly stripped of whitespace.

Daemon 1.59.2

Daemon 1.59 is polling-incompatible with previous daemon versions. Polling disruption will be minimal if all boxes in a PLN are updated to 1.59.2 within a few days of each other.

- Features
  - The Daemon Status/Archival Units page now comes up quickly. by omitting the Content Size and Disk Usage columns. Clicking the Show Sizes link will redisplay the page with the size columns.
  - The title set selection page (Journal Configuration/Add Titles) now comes up quickly.
  - The complete set of response headers received from the server with each file is available as Show All on the upper frame of the single URL status page (ViewContent).
  - The Title List report now allows filtering for journals only, books only, or all content.
  - Plugins may supply additional headers to be added to each HTTP request sent to the publisher (e.g., to add an Accept-Language: header to force the server to serve content to all boxes in a consistent language), by setting `au_http_request_header` to a string or list of strings of the form "*hdr*:*val*".
  - It's now possible to tailor substance patterns for AUs that are expected to contain only abstracts or tables of contents. If the title attribute `au_coverage_depth` is set (usually to FULLTEXT, ABSTRACT or TOC), and the plugin's `au_non_substance_url_pattern` or `au_substance_url_pattern` is a map, the `au_coverage_depth` value will be used as a key into that map.
  - Plugin-supplied URL normalizers may now alter any part of the URL. Previously they were not allowed to change the URL scheme, host or port. Set org.lockss.UrlUtil.unrestrictedSiteNormalize to false to restore the old behavior.
  - By default, polls will not delete (mark as deleted) files that a majority of other boxes don't

have. This is allows polls to converge more quickly when new files are being added. Plugins may control this for their AUs by setting `plugin_delete_extra_files` to true or false. The global default can be set with org.lockss.poll.v3.deleteExtraFiles (default now false, was formerly true).
    - Auto sizing of the Java heap now works correctly for the 64-bit JVM (which has much high memory requirements than 32-bit). Any manually added setting for `LOCKSS_JAVA_HEAP` in `/etc/lockss/config.dat` can be removed.
    - The `hostconfig` script allows multiple admin access subnets (semicolon-separated),

- Experimental Features
  These features are “beta” status as they haven’t been extensively exercised. They may change somewhat before being enabled by default in an upcoming release.

    - The metadata database schema have been extensively revised to be more complete and more flexible.
    - Usage data can be collected, and COUNTER reports produced.

- Bug Fixes
    - ServeContent and the proxy sometimes failed to find locally cached pages due to variations in URL-encoding.
    - ServeContent didn’t properly rewrite `https:` links.
    - The HTML parser used by hash and crawl filters and the link rewriter (ServeContent) failed to parse some `&lt:script>`sections that are routinely accepted by browsers. It’s now more permissive. which
        - fixes some formatting problems with served content.
        - potentially changes hash values used in polling. To avoid false disagreements, the polling protocol minor version has been incremented. Daemon 1.59 won’t participate in polls called by pre-1.59 daemons, and vice-versa.
    - The disk space %Full value is now correct.
    - The second and successive pages of the FileVersions table was truncated.
    - Some daemon exits caused by mis-formatted PDF files have been eliminated.

Daemon 1.58.3

- Features
    - Deep recrawls can be triggered from DebugPanel if org.lockss.debugPanel.deepCrawlEnabled=true .
    - The crawl refetch depth and actual link depth encountered are displayed in the crawl status.
    - A quick crawl test can be enabled either by setting the AU config parameter crawl_test_substance_threshold, when the AU is created, or by setting the title attribute of

same name, at any time, to the number of “substance” URLs that must be collected to consider the test a success.
    - URLs with inconclusive (“Too close”) tallies in polls can trigger a refetch from the publisher if either org.lockss.poll.v3.repairFromPublisherWhenTooClose is true, or the plugin sets plugin_repair_from_publisher_when_too_close to true.
    - The poll detail table includes more info about individual voters’ tallies.
    - Crawls can be stopped promptly by setting their priority (in org.lockss.crawler.crawlPriorityAuidMap) to -20000.
    - Metadata indexing can be prevented/aborted by setting AU’s priority to -10000/-20000 in org.lockss.metadataManager.indexPriorityAuidMap .
    - Title List reports can be customised to include a constant-valued column, by just adding a non-field-name value to the custom ordering. It no longer needs to be quoted, though it may be. A column will be included in which each cell contains the value.
    - CSV/TSV reports can be output directly from the Title List HTML output screen after inspection.
    - The header can be omitted from Title List reports.
- Bug fixes
    - ServeContent sometimes erroneously returned 404 for files that belong to multiple AUs but are not yet collected in all of them. This usually manifested as rendering problems due to missing css or image files.
    - ServeContent and the proxy could experience delays or timeouts when one or more browsers fetch multiple pages concurrently, due to insufficient default connection pool sizes.
    - Memory requirements for filtering large HTML files have been reduced.
    - Closed crawl windows no longer prevent DebugPanel/Start Crawl from enqueuing a crawl. The crawl will run after the window opens.
    - Title List can now produce a fully-compliant SFX format report, including the upper case “ACTIVE” column.
    - Titles lacking title_id values are excluded from SFX output.
- Known Problems
    - The %Full value in Repository Space displays is incorrect (low). It’s calculated as the free space / total space, instead of free space / usable space.

Daemon 1.57.4

- Features
    - Plugins may specify cookies to be sent with requests by setting au_http_cookie to a string or list of strings of the form *key=value*
- Bug fixes
    - Pages served by ServeContent sometimes had multiple, conflicting Content-Type headers, causing rendering problems..

- Manually queued polls start promptly.
- Restored display of average agreement to poll detail.
- Fixed a file descriptor leak in DebugPanel/Find Preserved URL.

Daemon 1.57.1

- Features
  - Title List can now limit the titles listed to just journals or just books. This allows separate reports to be generated whose print_identifier and online_identifier fields are just ISSNs or ISBNs.
  - The Title List field ordering table can now include a literal string of the form “FIELD-VALUE”. Each cell in the column, including the column title, will have the value without quotes. This feature allows formatting a table where the column must contain a fixed value. An example is the DataLoader for a local SFX link resolver instance. This avoids having to post-process the output report.
  - Specifying an OpenURL for a journal title now displays a page with information about the journal and a list of AUs for the journal that are preserved by the LOCKSS box. The information includes the journal name, publisher name, ISSN/eISSN, and a link to the journal if it is still available at the publisher.
  - Metadata indexing progress is displayed in Daemon Status.
  - Bulk content plugins (collections containing multiple titles/volumes) can/should set plugin_bulk_content to true. This is used (e.g., by MetadataManager) to determine when an AU’s Tdb entry doesn’t describe the actual contents. For now, this is automatically true of plugins whose name ends with “SourcePlugin”.
- Bug fixes
  - org.lockss.ui.hostNameInTitle did not take effect on Daemon Status pages.
  - AuState files were unnecessarily updated at startup, taking significant time when a large number of AUs are present.
  - Titlesets with names containing non-ascii characters did not work correctly in Add Titles, etc. forms.
  - OpenUrlResolver and ServeContent now respect the AU’s crawl proxy setting when accessing publisher content.
  - The metadata database is no longer locked for long periods of time during indexing.
  - Server timeouts while fetching permission pages could erroneously trigger thread watchdogs, causing daemon restarts.
  - The archive member separator (!/) is treated specially only if the plugin specifies plugin_archive_file_types, so can once again appear in URLs in normal AUs.
  - Some TAR files weren’t recognized as archives.
  - The UI can now be accessed via an IPv6 address containing a zone index. (Usually this happens with “localhost”.)

- The remaining bugs causing some AUs to fail some operations after their plugin is updated have been fixed.
- Title DB files packaged in plugin jars were not handled correctly when the plugin was updated. This feature has been disabled for now.

Daemon 1.56.3

- Features
  - Referrer URLs (the list of pages containing links to a particular page encountered in the crawl) can be displayed in crawl status tables. Set org.lockss.crawlStatus.recordReferrers = All (before the crawl starts). Lists of URLs in crawl status tables (fetched, excluded, etc.) will then have a Show Referrers link.
  - The list of AUs that contain content for a specified URL can displayed using the "Find Preserved URL" button in DebugPanel.
  - If org.lockss.ui.displayIpAddr = true, the machine's IP address will be displayed in the header of UI pages.
  - If org.lockss.ui.hostNameInTitle = true, page titles in the UI will start with the machine's hostname. Useful if you have lots of tabs open to different LOCKSS boxes.
  - The date/time format used in XML status table output (non-raw, *ie* without `outputVersion=2`) can be set with org.lockss.admin.xmlDataFormat, which should be set to `Local`, for local time, `GMT` or `UTC` for GMT, or a legal DateFormat string.
  - The number of bytes read and hashed by pollers and voters will be counted and displayed in poller and voter status pages if org.lockss.poll.v3.enableHashStats = true.
  - The crawler now follows `src` links in `<iframe>` tags. (*Ie*, the default link extractor extracts those URLs.)
  - A new PDF framework has been implemented, which greatly simplifies writing PDF filters for plugins.
  - Plugins can perform custom per-URL substance checking by setting `plugin_substance_predicate_factory` to a SubstancePredicateFactory implementation, whose makeSubstancePredicate(ArchivalUnit) returns a SubstancePredicate, whose isSubstanceUrl(String) method is called for each URL in the AU.
  - Polls started manually from DebugPanel are now queued and run when resources permit. Previously it was possible to start too many polls, causing resource exhaustion. (This does not occur during normal daemon operation.) Manual polls now require the poller to be enabled (org.lockss.poll.v3.enableV3Poller = true). To prevent other AUs from being polled, set org.lockss.poll.autoPollAuClassess = `Priority`.
  - org.lockss.poll.autoPollAuClassess determines which AUs may have polls started automatically. One or more of `All`, `Internal` (plugin repositories), `Priority` (started from DebugPanel).
  - AUs can be put through a quick crawl test which verifies that the manifest page can be

collected and has permission, and the AU contains at least N substance pages. Enable by setting the AU config param `crawl_test_substance_threshold` to the minimum number of substance pages required. If zero, only the manifest page is checked. The crawl stops with the status “Crawl test successful” as soon as the condition is met, or with “Crawl test failed” if the crawl ends successfully but the AU contains too few substance pages, or with another failure (*eg* “No permission”) if it doesn’t get that far.

- A crawl is scheduled for all plugin registries when org.lockss.plugin.crawlRegistriesOnce transitions from false to true.
- Database set up has been moved out of MetadataManager into a new DbManager. All configuration parameters of the form org.lockss.daemon.metadataManager.datasource.* should be changed to org.lockss.dbManager.datasource.*. (E.g., org.lockss.dbManager.datasource.className.

- Bug fixes
  - Agreement hint of 0.0 was incorrectly displayed in voter status when none was known
  - ServeContent handles several page-not-found situations more gracefully
  - The Reactivate AUs page no longer contains duplicate AUs.
  - KBART output better fits the recommendations. In particular, slightly disjoint ranges in coverage_notes fields are combined as necessary to fit within the maximum 500 characters.
  - A substantial reduction in the amount of memory consumed by large title databases.

- Developer-related
  - The `jvmargs` Ant property may be used to pass args to Java. *Eg*, to override the default java heap size for Ant builds and tests, use `-Djvmargs=-Xmx1024m` on the command like or add the line
    `jvmargs=-Xmx1024m`
    to `~/.lockssprops` . This may be necessary on 64-bit systems.
  - Per-user build settings should now go in `~/.lockssprops` . `~/lockss.properties` will continue to work.
  - `ant clean` now leaves generated tdbxml files alone. Use `clean-all` to delete them.

Daemon 1.55.3

See also the 1.54 changes below, which had not previously been released.

- Features
  - A generic metadata extractor for RIS citation files is available. RisMetadataExtractor reads line by line and maps 2 character RIS tags to metadata fields.
  - Child plugins may override an inherited value with “unspecified” (system default) by specifying the value `<org.lockss.util.Default />`
  - AU config params of type YEAR may now have the value 0.

    - The Title List can produce output with a single entry per title, with detailed coverage ranges appearing in the coverage_notes field.
    - Title List coverage_notes can be generated in SFX DataLoader format.

- Bug fixes
    - Crawl rate limits are followed more strictly across multiple crawls from the same publisher.
    - Some charset encoding changes in HTMLtags previously caused parsing failures in link rewriters, hash filters, etc. Now, offsets of up to org.lockss.filter.html.encodingMatchRange bytes (default 100) between the pre- and post- change character streams can be accomodated.
    - Loading updated plugins could leave some AUs in an inconsistent state, causing errors in the UI and other system processes until daemon restart. As a result of this fix, AUs can no longer be added by manually including their configuration in the global config file. Set org.lockss.plugin.allowGlobalAuConfig = true to restore the previous behavior (as well as the bug).
    - Restarting AUs when loading updated plugins needlessly caused the repository to be rescanned, causing the restart to take excessive time.

- UI Changes
    - The Plugins table includes links to each plugin's AUs.
    - Enabling UI features that require daemon restart (e.g., user accounting) displays a more helpful message if accessed before restart.
    - Changed "titles" to "AUs" in the title set selection screen and the AU selection screen (but not the Journal Configuration menu screen).

Daemon 1.54.3

Daemon 1.54 was released only to CLOCKSS machines for metadata collection experiments. All these changes are included in the 1.55.3 release.

- Features
    - The OpenURL resolver is able to locate a closer match for the requested bibliographic entity in many situations.
    - Individual members of preserved archive files (zip, tar, etc.) can be accessed using URLs of the form <url>!/<member>. Member names may include nested archive files as intermediate components (<url>!/<member>!/<member-of-member>). Plugins should set `plugin_archive_file_types` to `Standard` to have all known archive types recognized, or to a serialized instance of ArchiveFileTypes for finer control. CachedUrlSet iterators (e.g., for metadata extraction) can be instructed to include

archive members in the iteration, instead of the archive file itself. See SubTreeArticleIterator.Spec.setVisitArchiveMembers(boolean)
    - Crawl rate can be varied by time of day, day of week, etc. Plugins should specify a serialized RateLimiterInfo as the value of `au_rate_limiter_info`.
    - CrawlWindows.Daily is a simpler way for plugins to specify simple windows. CrawlWindows.Always and CrawlWindows.Never are always-open and never-open windows.
    - SubTreeArticleIterators can efficiently prune subtrees with:
      `setExcludeSubTreePattern(Pattern regex)`
      `setExcludeSubTreePatternTemplate(String patternTemplate)`
      `setIncludeSubTreePattern(Pattern regex)`
      `setIncludeSubTreePatternTemplate(String patternTemplate)`
    - Plugins may use the wildcard MIME type "*/*" to specify a default in any of MIME-type map. (Link extractors, Crawl Rate Limiters, etc.)
    - If a daemon-wide crawl proxy is in effect, the crawl status page displays it.
    - To specify a proxy for plugin registry AU crawls, set org.lockss.plugin.registries.crawlProxy to <host>:<port>, or to DIRECT to cancel a global proxy.
    - The DateFormat used for log message timestamps can be configured with org.lockss.log.timeStampFormat. E.g., set to "yyyy/MM/dd HH:mm:ss.SSS" to include full date and time on each entry.
    - Plugins should now specify permission-only URLs as the value of `au_permission_url`. The old name, `au_manifest`, will continue to work.
    - Re-enabled TSV (tab-separated values) output for the Title List.
- Bug fixes
    - A couple file descriptor leaks were fixed.
    - The Article List link on AU status page no longer requires plugin to have a MetadataExtractor, just an ArticleIterator.
    - URLs that redirect to permission pages, and probe permission pages and any redirects to them, were fetched but not stored in the AU (unless otherwise encountered in the crawl). Another daemon attempting to crawl this copy of the AU might fail due to 404s while following permission or probe redirections. If org.lockss.crawler.storePermissionScheme is set to StoreAllInSpec, the permission pages and all redirects will be stored, but they will also be subject to the crawl rules. If set to Legacy (the default), all same-host redirects are followed regardless of the crawl rules.
    - Some situations in which the content server served publisher error pages or generic indices, even though correct preserved content was available, have been fixed.
    - "X-Lockss-Result: Please" in UI requests once again returns the proper "X-Lockss-Result: Ok" (or "Fail") in the response.

- UI Changes
  - Poll agreement values are now displayed consistently across DaemonStatus tables. Some scripts may need to be adjusted. For normal output the value is a float between 0.0 and 100.0, followed by “%”. In XML output, raw agreement values (outputVersion=2) are an unadorned float between 0.0 and 1.0.