Appendix B2: Managing Heritrix 3 Crawler-Beans

A Heritrix3 harvest is defined by a Crawler-Bean (.cxml) file. This is a bean-definition file from the Spring framework. You can use Heritrix3's own documentation to create Crawler-Bean files which can then be uploaded to NetarchiveSuite via the GUI. NetarchiveSuite overwrites certain placeholder values in every Crawler-Bean definition before scheduling the harvest. The following placeholders are defined - some are required in every Crawler-Bean file, others are optional. When an optional placeholder is missing from the Crawler-Bean definition, then any attempt to redefine its value via the GUI will be ignored. There is no validation of Crawler-Bean files in this version of NetarchiveSuite, so a missing required placeholder will first manifest itself as a harvest job which fails to start. Some form for validation will be introduced in a later version of NetarchiveSuite.

Required Placeholders

Placeholder	Placing	Comments
`frontier.queueTotalBudget=%{FRONTIER_QUEUE_TOTAL_BUDGET_PLACEHOLDER}`	In PropertyOverrideConfigurer	See discussion below
`quotaenforcer.groupMaxFetchSuccesses=%{QUOTA_ENFORCER_GROUP_MAX_FETCH_SUCCES_PLACEHOLDER}`	In PropertyOverrideConfigurer	See discussion below
`quotaenforcer.groupMaxAllKb=%{QUOTA_ENFORCER_MAX_BYTES_PLACEHOLDER}`	In PropertyOverrideConfigurer	See discussion below
`%{CRAWLERTRAPS_PLACEHOLDER}`	in the regexList in MatchesListRegexDecideRule	Substituted with global crawler traps defined in NAS
`%{ARCHIVER_PROCESSOR_BEAN_PLACEHOLDER}`	At the first xml nesting level, inside the <beans> element	
`%{ARCHIVER_BEAN_REFERENCE_PLACEHOLDER}`	Inside the DispositionChain bean.	

Optional Placeholders

Placeholder	Placing	Comments
`crawlLimiter.maxTimeSeconds=%{MAX_TIME_SECONDS_PLACEHOLDER}`	In PropertyOverrideConfigurer	if absent, e.g. if maxTimeSeconds is hardcoded in the crawler-beans file, then NAS will never override this value.
`<property name="indexLocation" value="%{DEDUPLICATION_INDEX_LOCATION_PLACEHOLDER}"/>`	Inside the bean with class is.hi.bok.deduplicator.DeDuplicator	If absent, there will be no deduplication

`metadata.robotsPolicyName=%{HONOR_ROBOTS_DOT_TXT}`
or
`<property name="robotsPolicyName" value="%{HONOR_ROBOTS_DOT_TXT}"/>`
	In PropertyOverrideConfigurer
	or
	In metadata bean
		If absent, the robotsPolicy will be "ignore" (the default in H3) or hardwired to either obey or ignore

extractorHtml.extractJavascript=%{EXTRACT_JAVASCRIPT}	In PropertyOverrideConfigurer	If absent, the H3 template will use default value(?) or be hardwired to either true or false

scope.rules[2].maxHops=%{MAX_HOPS} (assuming TooManyHopsDecideRule is the 3rd bean defined in the "scope" bean) or `<property name="maxHops" value="%{MAX_HOPS}" />`	In PropertyOverrideConfigurer in bean for class `org.archive.modules.deciderules.TooManyHopsDecideRule`	If absent, the H3 template will use default value (20) or be hardwired to something else

Quote Enforcement

All three Quota/Budget -related placeholders are required, but their interpretation depends on the NAS setting `harvester.scheduler.jobGen.objectLimitIsSetByQuotaEnforcer`.

Behaviour is as follows:

objectLimitIsSetByQuotaEnforcer	
true	queueTotalBudget is set to infinity groupMaxFetchSuccesses is set to the maxObjectsPerDomain value from NAS
false	queueTotalBudget is set to the maxObjectsPerDomain value from NAS groupMaxFetchSuccesses is set to infinity

In all cases, groupMaxAllKb is set to the value determined from the maxBytesPerDomain setting from the NAS GUI (default value is -1 which is equivalent to no limit).