

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

Optional Placeholders

Placeholder	Placing	Comments
<code>crawlLimiter.maxTimeSeconds={MAX_TIME_SECONDS_PLACEHOLDER}</code>	In PropertyOverrideConfigurer	if absent, e.g. if maxTimeSeconds is hardcoded in the crawler-beans file, then NAS will never override this value.
<code><property name="indexLocation" value="{DEDUPLICATION_INDEX_LOCATION_PLACEHOLDER}" /></code>	Inside the bean with class is.hi.bok.deduplicator.DeDuplicator	If absent, there will be no deduplication
<code>metadata.robotsPolicyName={HONOR_ROBOTS_DOT_TXT}</code> or <code><property name="robotsPolicyName" value="{HONOR_ROBOTS_DOT_TXT}" /></code>	In PropertyOverrideConfigurer or In metadata bean	If absent, the robotsPolicy will be "ignore" (the default in H3) or hardwired to either obey or ignore
<code>extractorHtml.extractJavascript={EXTRACT_JAVASCRIPT}</code>	In PropertyOverrideConfigurer	If absent, the H3 template will use default value(?) or be hardwired to either true or false

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

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	<p>queueTotalBudget is set to infinity</p> <p>groupMaxFetchSuccesses is set to the maxObjectsPerDomain value from NAS</p>
false	<p>queueTotalBudget is set to the maxObjectsPerDomain value from NAS</p> <p>groupMaxFetchSuccesses is set to infinity</p>

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).