Migrating H1 templates to H3 to use with NetarchiveSuite 5.1+

General comments about H3 template in comparison with H1

- **It is perfectly possible to migrate from H1 to H3 simply by replacing each H1 template by an H3 template (with placeholders) using the "Edit Harvest Templates" section of the GUI.**
- The NewObject's in H1 have in most cases been turned into a Bean
- There is no longer a Scope class in H3, as the only type of scope in H3 is the DecidingScope. The bean called 'scope' is a DecideRuleSequence.
- The politeness settings are now combined in the 'disposition' bean.
- The properties of most beans can be added to the OVERRIDES section of the template.
  It is thus easier to manage the differences between templates.
- The `ContentSizeAnnotationPostProcessor` is still needed in H3:

```
<newObject name="ContentSize"
class="dk.netarkivet.harvester.harvesting.ContentSizeAnnotationPostPr
ocessor">
         <boolean name="enabled">true</boolean>
         <newObject name="ContentSize#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
         <map name="rules"> </map>
</newObject>
```

Otherwise the #bytes for each domain is not in the domainStats returned by the harvester. The following line

```
<bean id="ContentSizeAnnotationPostProcessor"
class="dk.netarkivet.harvester.harvesting.ContentSizeAnnotationPostProcessor"/>
```

needs to be inserted before the lines

<!-- ...send each outlink candidate URI to CandidatesChain,
and enqueue those ACCEPTed to the frontier... -->

<ref bean="candidates"/>

Placeholders

Required and option placeholders for crawler-bean templates are documented at **Appendix B2: Managing Heritrix 3 Crawler-Beans**

Basic H1- template for NetarchiveSuite use

```
<?xml version="1.0" encoding="UTF-8"?>
<crawl-order xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="heritrix_settings.xsd">
    <meta>
        <name>default_orderxml</name>
        <description>Default Profile</description>
        <operator>Admin</operator>
        <organization/>
        <audience/>
        <date>20080118111217</date>
    </meta>
    <controller>
        <string name="settings-directory">settings</string>
        <string name="disk-path"/>
        <string name="logs-path">logs</string>
        <string name="checkpoints-path">checkpoints</string>
        <string name="state-path">state</string>
```

```
        <string name="scratch-path">scratch</string>
        <long name="max-bytes-download">0</long>
        <long name="max-document-download">0</long>
        <long name="max-time-sec">0</long>
        <integer name="max-toe-threads">50</integer>
        <integer name="recorder-out-buffer-bytes">4096</integer>
        <integer name="recorder-in-buffer-bytes">65536</integer>
        <integer name="bdb-cache-percent">40</integer>
        <!-- DecidingScope migrated from DomainScope -->
        <newObject name="scope"
class="org.archive.crawler.deciderules.DecidingScope">
            <boolean name="enabled">true</boolean>
            <string name="seedsfile">seeds.txt</string>
            <boolean name="reread-seeds-on-config">true</boolean>
            <!-- DecideRuleSequence. Multiple DecideRules applied in order
with last non-PASS the resulting decision -->
            <newObject name="decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">

                <map name="rules">
                    <newObject name="rejectByDefault"
class="org.archive.crawler.deciderules.RejectDecideRule"/>
                    <newObject name="acceptURIFromSeedDomains"
class="dk.netarkivet.harvester.harvesting.OnNSDomainsDecideRule">
                        <string name="decision">ACCEPT</string>
                        <string name="surts-source-file"/>
                        <boolean
name="seeds-as-surt-prefixes">true</boolean>
                        <string name="surts-dump-file"/>
                        <boolean name="also-check-via">false</boolean>
                        <boolean name="rebuild-on-reconfig">true</boolean>
                    </newObject>
                    <newObject name="rejectIfTooManyHops"
class="org.archive.crawler.deciderules.TooManyHopsDecideRule">
                        <integer name="max-hops">25</integer>
                    </newObject>
                    <newObject name="rejectIfPathological"
class="org.archive.crawler.deciderules.PathologicalPathDecideRule">
                        <integer name="max-repetitions">3</integer>
                    </newObject>
                    <newObject name="acceptIfTranscluded"
class="org.archive.crawler.deciderules.TransclusionDecideRule">
                        <integer name="max-trans-hops">5</integer>
                        <integer name="max-speculative-hops">1</integer>
                    </newObject>
                    <newObject name="pathdepthfilter"
class="org.archive.crawler.deciderules.TooManyPathSegmentsDecideRule">
                        <integer name="max-path-depth">20</integer>
                    </newObject>
                    <newObject name="acceptIfPrerequisite"
class="org.archive.crawler.deciderules.PrerequisiteAcceptDecideRule">
                    </newObject>
                    <newObject name="globale_crawlertraps"
```

```
class="org.archive.crawler.deciderules.MatchesListRegExpDecideRule">
                        <string name="decision">REJECT</string>
                        <string name="list-logic">OR</string>
                        <stringList name="regexp-list"> <!-- a lot of
crawlertraps strings here have been removed for clarity -->
<string>.*twitter\.com.*(rss|logged|time.*\d\d:\d\d:\d\d).*</string>
<string>.*\/earch\/.*\/earch\/.*</string>
<string>.*ddc\.dk.*campaignmonitor.*campaignmonitor.*</string>
<string>.*thumbshots\.com.*url=[a-zA-Z0-9-]{1,}\.[a-z]{2,3}$.*</string>
<string>.*css.*css.*css(\w|\/\w|\.).*</string>

<string>.*scielo\.org\.ve.*</string>
<string>.*scielo\.sld\.cu.*</string>
<string>.*sciencedirect\.com.*</string>
<string>.*search\.ebscohost\.com.*</string>
<string>.*search\.epnet\.com.*</string>
<string>.*siam\.org.*</string>
<string>.*springerlink\.com.*</string>
<string>.*taylorandfrancis\.metapress\.com.*</string>
<string>.*thieme-connect\.com.*</string>
<string>.*worldscinet\.com.*</string>
<string>.*www3\.interscience\.wiley\.com.*</string>
<string>.*www-gdz\.sub\.uni-goettingen\.de.*</string>
<string>.*tlg\.uci\.edu.*</string>
                        </stringList>
<!-- end list of crawletraps -->
                    </newObject>
                </map> <!-- end rules -->
            </newObject> <!-- end decide-rules -->
        </newObject> <!-- End DecidingScope -->
        <map name="http-headers">
            <string name="user-agent">Mozilla/5.0 (compatible;
heritrix/1.14.4 +http://netarkivet.dk/webcrawler/)</string>
            <string name="from">info@netarkivet.dk</string>
        </map>
        <newObject name="robots-honoring-policy"
class="org.archive.crawler.datamodel.RobotsHonoringPolicy">
            <string name="type">ignore</string>
            <boolean name="masquerade">false</boolean>
            <text name="custom-robots"/>
            <stringList name="user-agents">
            </stringList>
        </newObject>
        <newObject name="frontier"
class="org.archive.crawler.frontier.BdbFrontier">
            <float name="delay-factor">1.0</float>
            <integer name="max-delay-ms">1000</integer>
            <integer name="min-delay-ms">300</integer>
            <integer name="max-retries">3</integer>
            <long name="retry-delay-seconds">300</long>
            <integer name="preference-embed-hops">1</integer>
            <integer name="total-bandwidth-usage-KB-sec">3000</integer>
            <integer
```

```
name="max-per-host-bandwidth-usage-KB-sec">500</integer>
            <string
name="queue-assignment-policy">dk.netarkivet.harvester.harvesting.Domainna
meQueueAssignmentPolicy</string>

            <string name="force-queue-assignment"/>
            <boolean name="pause-at-start">false</boolean>
            <boolean name="pause-at-finish">false</boolean>
            <boolean name="source-tag-seeds">false</boolean>
            <boolean name="recovery-log-enabled">false</boolean>
            <boolean name="hold-queues">true</boolean>
            <integer name="balance-replenish-amount">3000</integer>
            <integer name="error-penalty-amount">100</integer>
            <long name="queue-total-budget">-1</long>
            <string
name="cost-policy">org.archive.crawler.frontier.UnitCostAssignmentPolicy</
string>
            <long name="snooze-deactivate-ms">300000</long>
            <integer name="target-ready-backlog">50</integer>
            <string
name="uri-included-structure">org.archive.crawler.util.BdbUriUniqFilter</s
tring>
        </newObject>

        <map name="uri-canonicalization-rules">
            <newObject name="stillinger_i_staten_timecode"
class="org.archive.crawler.url.canonicalize.RegexRule">
                <boolean name="enabled">true</boolean>
                <string
name="matching-regex">^(.*stillinger-i-staten.dk.*)(tc=.*)$</string>
                <string name="format">${1}</string>
                <string name="comment">fjerner tc=... fra
stillinger-i-staten.dk</string>
            </newObject>
            <newObject name="Lowercase"
class="org.archive.crawler.url.canonicalize.LowercaseRule">
                <boolean name="enabled">true</boolean>
            </newObject>
            <newObject name="Userinfo"
class="org.archive.crawler.url.canonicalize.StripUserinfoRule">
                <boolean name="enabled">true</boolean>
            </newObject>
            <newObject name="WWW"
class="org.archive.crawler.url.canonicalize.StripWWWRule">
                <boolean name="enabled">false</boolean>
            </newObject>
            <newObject name="SessionIDs"
class="org.archive.crawler.url.canonicalize.StripSessionIDs">
                <boolean name="enabled">true</boolean>
            </newObject>
            <newObject name="QueryStrPrefix"
class="org.archive.crawler.url.canonicalize.FixupQueryStr">
                <boolean name="enabled">true</boolean>
```

```
            </newObject>
        </map>
        <!-- Heritrix pre-fetch processors -->
        <map name="pre-fetch-processors">

            <newObject name="QuotaEnforcer"
class="org.archive.crawler.prefetch.QuotaEnforcer">
                <boolean name="force-retire">false</boolean>
                <boolean name="enabled">true</boolean>
                <newObject name="QuotaEnforcer#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
                <long name="server-max-fetch-successes">-1</long>
                <long name="server-max-success-kb">-1</long>
                <long name="server-max-fetch-responses">-1</long>
                <long name="server-max-all-kb">-1</long>

                <long name="host-max-fetch-successes">-1</long>
                <long name="host-max-success-kb">-1</long>
                <long name="host-max-fetch-responses">-1</long>
                <long name="host-max-all-kb">-1</long>

                <long name="group-max-fetch-successes">-1</long>
                <long name="group-max-success-kb">-1</long>
                <long name="group-max-fetch-responses">-1</long>
                <long name="group-max-all-kb">-1</long>

            </newObject>

            <newObject name="Preselector"
class="org.archive.crawler.prefetch.Preselector">
                <boolean name="enabled">true</boolean>
                <newObject name="Preselector#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
                <boolean name="override-logger">false</boolean>
                <boolean name="recheck-scope">true</boolean>
                <boolean name="block-all">false</boolean>
                <string name="block-by-regexp"/>
                <string name="allow-by-regexp"/>
            </newObject>
            <newObject name="Preprocessor"
class="org.archive.crawler.prefetch.PreconditionEnforcer">
                <boolean name="enabled">true</boolean>
                <newObject name="Preprocessor#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
```

```
                <integer
name="ip-validity-duration-seconds">21600</integer>
                <integer
name="robot-validity-duration-seconds">86400</integer>
                <boolean name="calculate-robots-only">false</boolean>
            </newObject>
        </map>
        <!--End of Heritrix pre-fetch processors -->
        <!-- Heritrix fetch processors -->
        <map name="fetch-processors">
            <newObject name="DNS"
class="org.archive.crawler.fetcher.FetchDNS">
                <boolean name="enabled">true</boolean>
                <newObject name="DNS#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
                <boolean name="accept-non-dns-resolves">false</boolean>
                <boolean name="digest-content">true</boolean>
                <string name="digest-algorithm">sha1</string>

            </newObject>
            <newObject name="HTTP"
class="org.archive.crawler.fetcher.FetchHTTP">
                <boolean name="enabled">true</boolean>
                <newObject name="HTTP#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
                <newObject name="midfetch-decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
                <integer name="timeout-seconds">1200</integer>
                <integer name="sotimeout-ms">20000</integer>
                <integer name="fetch-bandwidth">0</integer>
                <long name="max-length-bytes">0</long>
                <boolean name="ignore-cookies">false</boolean>
                <boolean name="use-bdb-for-cookies">true</boolean>
                <string name="load-cookies-from-file"/>
                <string name="save-cookies-to-file"/>
                <string name="trust-level">open</string>
                <stringList name="accept-headers">
                </stringList>
                <string name="http-proxy-host"/>
                <string name="http-proxy-port"/>
                <string name="default-encoding">ISO-8859-1</string>
                <boolean name="digest-content">true</boolean>
                <string name="digest-algorithm">sha1</string>
                <boolean name="send-if-modified-since">true</boolean>
```

```
                <boolean name="send-if-none-match">true</boolean>
                <boolean name="send-connection-close">true</boolean>
                <boolean name="send-referer">true</boolean>
                <boolean name="send-range">false</boolean>
                <string name="http-bind-address"/>
            </newObject>
        </map> <!-- end of Heritrix Fetch processors -->

        <!-- Heritrix extract processors -->
        <map name="extract-processors">
            <newObject name="ExtractorHTTP"
class="org.archive.crawler.extractor.ExtractorHTTP">
                <boolean name="enabled">true</boolean>
                <newObject name="ExtractorHTTP#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
            </newObject>
            <newObject name="ExtractorHTML"
class="org.archive.crawler.extractor.ExtractorHTML">
                <boolean name="enabled">true</boolean>
                <newObject name="ExtractorHTML#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
                <boolean name="extract-javascript">true</boolean>
                <boolean name="treat-frames-as-embed-links">false</boolean>
                <boolean name="ignore-form-action-urls">true</boolean>
                <boolean name="extract-value-attributes">false</boolean>
                <boolean name="ignore-unexpected-html">true</boolean>
            </newObject>
            <newObject name="ExtractorCSS"
class="org.archive.crawler.extractor.ExtractorCSS">
                <boolean name="enabled">true</boolean>
                <newObject name="ExtractorCSS#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
            </newObject>
            <newObject name="ExtractorJS"
class="org.archive.crawler.extractor.ExtractorJS">
                <boolean name="enabled">true</boolean>
                <newObject name="ExtractorJS#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
            </newObject>
            <newObject name="ExtractorSWF"
class="org.archive.crawler.extractor.ExtractorSWF">
```

```
                <boolean name="enabled">true</boolean>
                <newObject name="ExtractorSWF#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
            </newObject>
            <newObject name="ExtractorImpliedURI1"
class="org.archive.crawler.extractor.ExtractorImpliedURI">
                <boolean name="enabled">true</boolean>
                <string
name="trigger-regexp">^(http://www.e-pages.dk/urban/[0-9]*/)$</string>
                <string name="build-pattern">$1print.pdf</string>
                <boolean name="remove-trigger-uris">false</boolean>
            </newObject>

        </map> <!-- end of Heritrix extract processors -->
        <!-- Heritrix write processors -->
        <map name="write-processors">
<!-- Deduplicator process -->
            <newObject name="DeDuplicator"
class="is.hi.bok.deduplicator.DeDuplicator">
                <boolean name="enabled">true</boolean>
                <map name="filters">
                </map>
                <string name="index-location"/>
                <string name="matching-method">By URL</string>
                <boolean name="try-equivalent">true</boolean>
                <boolean name="change-content-size">false</boolean>
                <string name="mime-filter">^text/.*</string>
                <string name="filter-mode">Blacklist</string>
                <string name="analysis-mode">Timestamp</string>
                <string name="log-level">SEVERE</string>
                <string name="origin"/>
                <string name="origin-handling">Use index
information</string>
                <boolean name="stats-per-host">true</boolean>
        <boolean name="use-sparse-range-filter">true</boolean>
            </newObject>
<newObject name="WARCArchiver"
class="dk.netarkivet.harvester.harvesting.WARCWriterProcessor">
<boolean name="enabled">true</boolean>
<newObject name="WARCArchiver#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
<map name="rules"/>
</newObject>
<boolean name="compress">false</boolean>
<string name="prefix">netarkivet</string>
<string name="suffix">${HOSTNAME}</string>
<long name="max-size-bytes">1000000000</long>
<stringList name="path"> <string>warcs</string> </stringList>
<integer name="pool-max-active">5</integer>
<integer name="pool-max-wait">300000</integer>
```

```
<long name="total-bytes-to-write">0</long>
<boolean name="skip-identical-digests">false</boolean>
<boolean name="write-requests">true</boolean>
<boolean name="write-metadata">true</boolean>
<boolean name="write-revisit-for-identical-digests">true</boolean>
<boolean name="write-revisit-for-not-modified">true</boolean>
</newObject>
        </map> <!-- End of Heritrix write processors -->
        <!-- Heritrix post processors -->
        <map name="post-processors">
            <newObject name="Updater"
class="org.archive.crawler.postprocessor.CrawlStateUpdater">
                <boolean name="enabled">true</boolean>
                <newObject name="Updater#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
            </newObject>
            <newObject name="LinksScoper"
class="org.archive.crawler.postprocessor.LinksScoper">
                <boolean name="enabled">true</boolean>
                <newObject name="LinksScoper#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
                <boolean name="override-logger">false</boolean>
                <boolean name="seed-redirects-new-seed">false</boolean>
                <integer name="preference-depth-hops">-1</integer>

                <newObject name="scope-rejected-url-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
            </newObject>

            <newObject name="Scheduler"
class="org.archive.crawler.postprocessor.FrontierScheduler">
                <boolean name="enabled">true</boolean>
                <newObject name="Scheduler#decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
            </newObject>

            <newObject name="ContentSize"
class="dk.netarkivet.harvester.harvesting.ContentSizeAnnotationPostProcess
or">
                <boolean name="enabled">true</boolean>
                <newObject name="ContentSize#decide-rules"
```

```
class="org.archive.crawler.deciderules.DecideRuleSequence">
                    <map name="rules">
                    </map>
                </newObject>
            </newObject>

        </map>  <!-- end of Heritrix post processors -->

        <map name="loggers">
            <newObject name="crawl-statistics"
class="org.archive.crawler.admin.StatisticsTracker">
                <integer name="interval-seconds">20</integer>
            </newObject>
        </map>
        <string name="recover-path"/>
        <boolean name="checkpoint-copy-bdbje-logs">true</boolean>
        <boolean name="recover-retain-failures">false</boolean>
<!--  credentials -->
        <newObject name="credential-store"
class="org.archive.crawler.datamodel.CredentialStore">
            <map name="credentials">
                <newObject name="licitationen_login_1"
class="org.archive.crawler.datamodel.credential.Rfc2617Credential">
                    <string
name="credential-domain">www.licitationen.dk</string>
                    <string name="realm">Dagbladet Licitationen</string>
                    <string name="login">*****</string>
                    <string name="password">******</string>
                </newObject>
                <newObject name="mymusic_login_1"
class="org.archive.crawler.datamodel.credential.HtmlFormCredential">
                    <string
name="credential-domain">www.mymusic.dk</string>
                    <string
name="login-uri">http://www.mymusic.dk/konto/login2.asp</string>
                    <string name="http-method">POST</string>
                    <map name="form-items">
                        <string name="username">*****</string>
                        <string name="password">*****</string>
                        <string name="autologin">y</string>
                    </map>
                </newObject>
                <newObject name="arto_login_1"
class="org.archive.crawler.datamodel.credential.HtmlFormCredential">
                    <string name="credential-domain">www.arto.dk</string>
                    <string
name="login-uri">http://www.arto.dk/r2/frames/navigation.asp</string>
                    <string name="http-method">POST</string>
                    <map name="form-items">
                        <string name="action">submit</string>
                        <string name="brugernavn">*****</string>
                        <string name="kodeord">*******</string>
                        <string name="AutoLogin">Ja</string>
```

```
                        <string name="loginKnap">Log ind</string>
                    </map>
                </newObject>
                <newObject name="Heerfordt.dk"
class="org.archive.crawler.datamodel.credential.HtmlFormCredential">
                    <string name="credential-domain">heerfordt.dk</string>
                    <string name="login-uri">http://heerfordt.dk/</string>
                    <string name="http-method">POST</string>
                    <map name="form-items">
                        <string name="Brugernavn">******</string>
                        <string name="Pw">*******</string>
                        <string name="Login">Login</string>
                    </map>
                </newObject>
            </map>
```

```
        </newObject>
    </controller>
</crawl-order>
```

H3 template matching the above H1 template w/ NAS necessary placeholders

```
<?xml version="1.0" encoding="UTF-8"?>
<!--
  HERITRIX 3 CRAWL JOB CONFIGURATION FILE - MIGRATE TEMPLATE

   This is a relatively minimal configuration suitable for many crawls.

   Commented-out beans and properties are provided as an example; values
   shown in comments reflect the actual defaults which are in effect
   without specification. (To change from the default behavior,
   uncomment AND alter the shown values.)

   This is also the first step towards a way of migrating our
NetarchiveSuite H1 templates to H3.3.0

   This means adding beans for a QuotaEnforcer, a DeDuplicator, a
WARCWriterProcessor with added WarcInfo metadata.

 -->
<beans xmlns="http://www.springframework.org/schema/beans"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:context="http://www.springframework.org/schema/context"
        xmlns:aop="http://www.springframework.org/schema/aop"
        xmlns:tx="http://www.springframework.org/schema/tx"
        xsi:schemaLocation="http://www.springframework.org/schema/beans
http://www.springframework.org/schema/beans/spring-beans-3.0.xsd
          http://www.springframework.org/schema/aop
http://www.springframework.org/schema/aop/spring-aop-3.0.xsd
          http://www.springframework.org/schema/tx
http://www.springframework.org/schema/tx/spring-tx-3.0.xsd
          http://www.springframework.org/schema/context
http://www.springframework.org/schema/context/spring-context-3.0.xsd">

 <context:annotation-config/>

<!--
  OVERRIDES
   Values elsewhere in the configuration may be replaced ('overridden')
   by a Properties map declared in a PropertiesOverrideConfigurer,
   using a dotted-bean-path to address individual bean properties.
   This allows us to collect a few of the most-often changed values
   in an easy-to-edit format here at the beginning of the model
   configuration.
 -->
 <!-- overrides from a text property list -->
 <bean id="simpleOverrides"
```

```
class="org.springframework.beans.factory.config.PropertyOverrideConfigurer
">
  <property name="properties">
<!-- Overrides the default values used by Heritrix -->
   <value>
# This Properties map is specified in the Java 'property list' text format
#
http://java.sun.com/javase/6/docs/api/java/util/Properties.html#load%28jav
a.io.Reader%29

###
### some of these overrides is actually just the default value, so they can
be skipped
###

## Q: can overrides like 'fetchDns.enabled=false' be used to disable the
beans?

metadata.jobName=default_orderxml
metadata.description=Default Profile
metadata.operator=Admin
metadata.userAgentTemplate=Mozilla/5.0 (compatible; heritrix/3.3.0
+@OPERATOR_CONTACT_URL@)
## Edit the two following lines to match your setup.
metadata.operatorContactUrl=http://netarkivet.dk/webcrawler/
metadata.operatorFrom=info@netarkivet.dk

loggerModule.path=logs

crawlLimiter.maxBytesDownload=0
crawlLimiter.maxDocumentsDownload=0
## MaxTimeseconds inserted by NetarchiveSuite (Delete line, if behaviour
unwanted)
crawlLimiter.maxTimeSeconds=%{MAX_TIME_SECONDS_PLACEHOLDER}

crawlController.maxToeThreads=50
crawlController.recorderOutBufferBytes=4096
crawlController.recorderInBufferBytes=65536
crawlController.pauseAtStart=false
crawlController.scratchDir=scratch

## org.archive.bdb.BdbModule overrides
bdb.dir=state
bdb.cachePercent=40

## seeds properties
seeds.sourceTagSeeds=false

scope.rules[2].maxHops=25
scope.rules[6].maxRepetitions=3
scope.rules[3].maxTransHops=5
scope.rules[3].maxSpeculativeHops=1
```

```
## Politeness overrides
disposition.delayFactor=1.0
disposition.maxDelayMs=1000
disposition.minDelayMs=300
disposition.maxPerHostBandwidthUsageKbSec=500

preparer.preferenceEmbedHops=1
preparer.preferenceDepthHops=-1

frontier.maxRetries=3
frontier.retryDelaySeconds=300
frontier.recoveryLogEnabled=false
frontier.balanceReplenishAmount=3000
frontier.errorPenaltyAmount=100

frontier.queueTotalBudget=%{FRONTIER_QUEUE_TOTAL_BUDGET_PLACEHOLDER}
frontier.snoozeLongMs=300000

preselector.enabled=true
preselector.logToFile=false
preselector.recheckScope=true
preselector.blockAll=false

preconditions.enabled=true
preconditions.ipValidityDurationSeconds=21600
preconditions.robotsValidityDurationSeconds=86400
preconditions.calculateRobotsOnly=false

fetchDns.enabled=true
fetchDns.acceptNonDnsResolves=false
fetchDns.digestContent=true
fetchDns.digestAlgorithm=sha1

fetchHttp.enabled=true
fetchHttp.timeoutSeconds=1200
fetchHttp.soTimeoutMs=20000
fetchHttp.maxFetchKBSec=0
fetchHttp.maxLengthBytes=0
fetchHttp.ignoreCookies=false
fetchHttp.sslTrustLevel=OPEN
fetchHttp.defaultEncoding=ISO-8859-1
fetchHttp.digestContent=true
fetchHttp.digestAlgorithm=sha1
fetchHttp.sendIfModifiedSince=true
fetchHttp.sendIfNoneMatch=true
fetchHttp.sendConnectionClose=true
fetchHttp.sendReferer=true
fetchHttp.sendRange=false
extractorHttp.enabled=true
extractorHtml.enabled=true
extractorHtml.extractJavascript=true
extractorHtml.treatFramesAsEmbedLinks=false
extractorHtml.ignoreFormActionUrls=true
```

```
extractorHtml.extractValueAttributes=false
extractorHtml.ignoreUnexpectedHtml=true
extractorCss.enabled=true
extractorJs.enabled=true
extractorSwf.enabled=true

candidates.seedsRedirectNewSeeds=false
statisticsTracker.intervalSeconds=20

   </value>
  </property>
 </bean>

 <!-- overrides from declared <prop> elements, more easily allowing
      multiline values or even declared beans -->
 <bean id="longerOverrides"
class="org.springframework.beans.factory.config.PropertyOverrideConfigurer
">
  <property name="properties">
   <props>
   </props>
  </property>
 </bean>

 <!-- CRAWL METADATA: including identification of crawler/operator -->
 <bean id="metadata" class="org.archive.modules.CrawlMetadata"
autowire="byName">
       <property name="operatorContactUrl" value="[see override above]"/>
       <property name="jobName" value="[see override above]"/>
       <property name="description" value="[see override above]"/>
       <property name="robotsPolicyName" value="ignore"/>
  <!-- <property name="operator" value=""/> -->
  <!-- <property name="operatorFrom" value=""/> -->
  <!-- <property name="organization" value=""/> -->
  <!-- <property name="audience" value=""/> -->
  <!-- <property name="userAgentTemplate"
         value="Mozilla/5.0 (compatible; heritrix/@VERSION@
+@OPERATOR_CONTACT_URL@)"/> -->

 </bean>

 <!-- SEEDS: crawl starting points -->
 <!-- ConfigFile approach: specifying external seeds.txt file -->
 <bean id="seeds" class="org.archive.modules.seeds.TextSeedModule">
  <property name="textSource">
   <bean class="org.archive.spring.ConfigFile">
    <property name="path" value="seeds.txt" />
   </bean>
  </property>
  <property name="sourceTagSeeds" value="false"/>
 </bean>

 <!-- SCOPE: rules for which discovered URIs to crawl; order is very
```

```
      important because last decision returned other than 'NONE' wins. -->
 <bean id="scope"
class="org.archive.modules.deciderules.DecideRuleSequence">
  <property name="rules">
   <list>
    <!-- Begin by REJECTing all... -->
    <bean class="org.archive.modules.deciderules.RejectDecideRule">
    </bean>
    <!-- ...then ACCEPT those within configured/seed-implied SURT
prefixes... -->
    <bean
class="org.archive.modules.deciderules.surt.SurtPrefixedDecideRule">
     <!-- <property name="seedsAsSurtPrefixes" value="true" /> -->
     <!-- <property name="alsoCheckVia" value="true" /> -->
     <!-- <property name="surtsSourceFile" value="" /> -->
     <!-- <property name="surtsDumpFile" value="surts.dump" /> -->
    </bean>
    <!-- ...but REJECT those more than a configured link-hop-count from
start... -->
    <bean class="org.archive.modules.deciderules.TooManyHopsDecideRule">
     <!-- <property name="maxHops" value="20" /> -->
    </bean>
    <!-- ...but ACCEPT those more than a configured link-hop-count from
start... -->
    <bean class="org.archive.modules.deciderules.TransclusionDecideRule">
     <!-- <property name="maxTransHops" value="2" /> -->
     <!-- <property name="maxSpeculativeHops" value="1" /> -->
    </bean>
    <!-- ...but REJECT those from a configurable (initially empty) set of
REJECT SURTs... -->
    <bean
class="org.archive.modules.deciderules.surt.SurtPrefixedDecideRule">
          <property name="decision" value="REJECT"/>
          <property name="seedsAsSurtPrefixes" value="false"/>
          <property name="surtsDumpFile" value="negative-surts.dump" />
     <!-- <property name="surtsSourceFile" value="" /> -->
    </bean>
    <!-- ...and REJECT those from a configurable (initially empty) set of
URI regexes... -->
    <bean
class="org.archive.modules.deciderules.MatchesListRegexDecideRule">
     <property name="listLogicalOr" value="true" />
     <property name="regexList">
           <list>




<value>.*twitter\.com.*(rss|logged|time.*\d\d:\d\d:\d\d).*</value>
<value>.*\/earch\/.*\/earch\/.*</value>
<value>.*ddc\.dk.*campaignmonitor.*campaignmonitor.*</value>
```

```
<value>.*thumbshots\.com.*url=[a-zA-Z0-9-]{1,}\.[a-z]{2,3}$.*</value>
<value>.*css.*css.*css(\w|\/\w|\.).*</value>
<value>.*scielo\.org\.ve.*</value>
<value>.*scielo\.sld\.cu.*</value>
<value>.*sciencedirect\.com.*</value>
<value>.*search\.ebscohost\.com.*</value>
<value>.*search\.epnet\.com.*</value>
<value>.*siam\.org.*</value>
<value>.*springerlink\.com.*</value>
<value>.*taylorandfrancis\.metapress\.com.*</value>
<value>.*thieme-connect\.com.*</value>
<value>.*worldscinet\.com.*</value>
<value>.*www3\.interscience\.wiley\.com.*</value>
<value>.*www-gdz\.sub\.uni-goettingen\.de.*</value>
<value>.*tlg\.uci\.edu.*</value>

<!-- Here we inject our global crawlertraps, domain specific crawlertraps
-->


%{CRAWLERTRAPS_PLACEHOLDER}
           </list>
          </property>
    </bean>

    <!-- ...and REJECT those with suspicious repeating path-segments... -->
    <bean
class="org.archive.modules.deciderules.PathologicalPathDecideRule">
     <!-- <property name="maxRepetitions" value="2" /> -->
    </bean>
    <!-- ...and REJECT those with more than threshold number of
path-segments... -->
    <bean
class="org.archive.modules.deciderules.TooManyPathSegmentsDecideRule">
     <!-- <property name="maxPathDepth" value="20" /> -->
    </bean>
    <!-- ...but always ACCEPT those marked as prerequisites for another
URI... -->
    <bean
class="org.archive.modules.deciderules.PrerequisiteAcceptDecideRule">
    </bean>
    <!-- ...but always REJECT those with unsupported URI schemes -->
    <bean class="org.archive.modules.deciderules.SchemeNotInSetDecideRule">
    </bean>
   </list>
  </property>
 </bean>

 <!--
   PROCESSING CHAINS
    Much of the crawler's work is specified by the sequential
    application of swappable Processor modules. These Processors
    are collected into three 'chains. The CandidateChain is applied
```

```
    to URIs being considered for inclusion, before a URI is enqueued
    for collection. The FetchChain is applied to URIs when their
    turn for collection comes up. The DispositionChain is applied
    after a URI is fetched and analyzed/link-extracted.
  -->

 <!-- CANDIDATE CHAIN -->
 <!-- processors declared as named beans -->
 <bean id="candidateScoper"
class="org.archive.crawler.prefetch.CandidateScoper">
 </bean>
 <bean id="preparer" class="org.archive.crawler.prefetch.FrontierPreparer">
  <!-- <property name="preferenceDepthHops" value="-1" /> -->
  <!-- <property name="preferenceEmbedHops" value="1" /> -->
  <!-- <property name="canonicalizationPolicy">
        <ref bean="canonicalizationPolicy" />
       </property> -->
   <property name="queueAssignmentPolicy">
        <ref bean="queueAssignmentPolicy" />
<!-- Bundled with NAS is two queueAssignPolicies (code is in
heritrix3-extensions):
 dk.netarkivet.harvester.harvesting.DomainnameQueueAssignmentPolicy
 dk.netarkivet.harvester.harvesting.SeedUriDomainnameQueueAssignmentPolicy
-->
       </property>

 <!-- <property name="uriPrecedencePolicy">
        <ref bean="uriPrecedencePolicy" />
       </property> -->
  <!-- <property name="costAssignmentPolicy">
        <ref bean="costAssignmentPolicy" />
       </property> -->
 </bean>
 <!-- assembled into ordered CandidateChain bean -->
 <bean id="candidateProcessors" class="org.archive.modules.CandidateChain">
  <property name="processors">
   <list>
    <!-- apply scoping rules to each individual candidate URI... -->
    <ref bean="candidateScoper"/>
    <!-- ...then prepare those ACCEPTed for enqueuing to frontier. -->
    <ref bean="preparer"/>
   </list>
  </property>
 </bean>

 <!-- FETCH CHAIN -->
 <!-- processors declared as named beans -->
 <bean id="preselector" class="org.archive.crawler.prefetch.Preselector">
  <!-- <property name="recheckScope" value="false" /> -->
  <!-- <property name="blockAll" value="false" /> -->
  <!-- <property name="blockByRegex" value="" /> -->
  <!-- <property name="allowByRegex" value="" /> -->
 </bean>
```

```
 <bean id="preconditions"
class="org.archive.crawler.prefetch.PreconditionEnforcer">

  <!-- refer to a list of credentials -->
  <property name="credentialStore">
    <ref bean="credentialStore" />
  </property>
 </bean>
 <bean id="fetchDns" class="org.archive.modules.fetcher.FetchDNS">
 </bean>
 <bean id="fetchHttp" class="org.archive.modules.fetcher.FetchHTTP">
 </bean>
 <bean id="extractorHttp"
class="org.archive.modules.extractor.ExtractorHTTP">
 </bean>
 <bean id="extractorHtml"
class="org.archive.modules.extractor.ExtractorHTML">
 </bean>
 <bean id="extractorCss"
class="org.archive.modules.extractor.ExtractorCSS">
 </bean>

 <bean id="extractorJs" class="org.archive.modules.extractor.ExtractorJS">
 </bean>

 <bean id="extractorSwf"
class="org.archive.modules.extractor.ExtractorSWF">
 </bean>

 <!-- assembled into ordered FetchChain bean -->
 <bean id="fetchProcessors" class="org.archive.modules.FetchChain">
  <property name="processors">
   <list>
    <!-- recheck scope, if so enabled... -->
    <ref bean="preselector"/>
    <!-- ...then verify or trigger prerequisite URIs fetched, allow
crawling... -->
    <ref bean="preconditions"/>

    <!-- check, if quotas is already superseded -->
    <ref bean="quotaenforcer"/>  <!-- always required by NAS ? -->

    <!-- ...fetch if DNS URI... -->
    <ref bean="fetchDns"/>
    <!-- ...fetch if HTTP URI... -->
    <ref bean="fetchHttp"/>
    <!-- ...extract oulinks from HTTP headers... -->
    <ref bean="extractorHttp"/>
    <!-- ...extract oulinks from HTML content... -->
    <ref bean="extractorHtml"/>
    <!-- ...extract oulinks from CSS content... -->
    <ref bean="extractorCss"/>
    <!-- ...extract oulinks from Javascript content... -->
```

```
    <ref bean="extractorJs"/>
    <!-- ...extract oulinks from Flash content... -->
    <ref bean="extractorSwf"/>
    <bean id="extractorEpagesDk"
class="org.archive.modules.extractor.ExtractorImpliedURI">
 <property name="regex" value="^(http://www.e-pages.dk/urban/[0-9]*/)$"/>
 <property name="format" value="$1print.pdf"/>
        <property name="removeTriggerUris" value="false"</>
    </bean>
   </list>
  </property>
 </bean>

 <!-- DISPOSITION CHAIN -->
 <!-- processors declared as named beans -->

<!-- Here the (W)arc writer is inserted -->
%{ARCHIVER_PROCESSOR_BEAN_PLACEHOLDER}

<bean id="DeDuplicator" class="is.hi.bok.deduplicator.DeDuplicator">
<!-- DEDUPLICATION_INDEX_LOCATION_PLACEHOLDER is replaced by path on
harvest-server -->
        <property name="indexLocation"
value="%{DEDUPLICATION_INDEX_LOCATION_PLACEHOLDER}"/>
        <property name="matchingMethod" value="URL"/>
        <property name="tryEquivalent" value="TRUE"/>
        <property name="changeContentSize" value="false"/>
        <property name="mimeFilter" value="^text/.*"/>
        <property name="filterMode" value="BLACKLIST"/>
<!--  <property name="analysisMode" value="TIMESTAMP"/> FIXME -->
        <property name="origin" value=""/>
        <property name="originHandling" value="INDEX"/>
        <property name="statsPerHost" value="true"/>
</bean>

 <bean id="candidates"
class="org.archive.crawler.postprocessor.CandidatesProcessor">
  <!-- <property name="seedsRedirectNewSeeds" value="true" /> -->
 </bean>
 <bean id="disposition"
class="org.archive.crawler.postprocessor.DispositionProcessor">
 </bean>

 <!-- assembled into ordered DispositionChain bean -->
 <bean id="dispositionProcessors"
class="org.archive.modules.DispositionChain">
  <property name="processors">
   <list>
    <!-- write to aggregate archival files... -->

    <!-- remove the reference below, and the DeDuplicator bean itself to
disable Deduplication -->
    <ref bean="DeDuplicator"/>
```

```
    <!-- Here the reference to the (w)arcWriter bean is inserted -->

    %{ARCHIVER_BEAN_REFERENCE_PLACEHOLDER}


    <!-- The next processor is required. otherwise domain bytecounts is not
reported back from the harvester -->
    <bean id="ContentSizeAnnotationPostProcessor"
class="dk.netarkivet.harvester.harvesting.ContentSizeAnnotationPostProcess
or"/>

    <!-- ...send each outlink candidate URI to CandidatesChain,
         and enqueue those ACCEPTed to the frontier... -->
    <ref bean="candidates"/>
    <!-- ...then update stats, shared-structures, frontier decisions -->
    <ref bean="disposition"/>
   </list>
  </property>
 </bean>

 <!-- CRAWLCONTROLLER: Control interface, unifying context -->
 <bean id="crawlController"
   class="org.archive.crawler.framework.CrawlController">
 </bean>

 <!-- FRONTIER: Record of all URIs discovered and queued-for-collection -->
 <bean id="frontier"
   class="org.archive.crawler.frontier.BdbFrontier">
 </bean>

 <!-- URI UNIQ FILTER: Used by frontier to remember already-included URIs
-->
 <bean id="uriUniqFilter"
   class="org.archive.crawler.util.BdbUriUniqFilter">
 </bean>

 <!--
   OPTIONAL BUT RECOMMENDED BEANS
  -->

 <!-- ACTIONDIRECTORY: disk directory for mid-crawl operations
      Running job will watch directory for new files with URIs,
      scripts, and other data to be processed during a crawl. -->
 <bean id="actionDirectory"
class="org.archive.crawler.framework.ActionDirectory">
 </bean>

 <!--  CRAWLLIMITENFORCER: stops crawl when it reaches configured limits
-->
 <bean id="crawlLimiter"
class="org.archive.crawler.framework.CrawlLimitEnforcer">
  </bean>
```

```
 <!-- CHECKPOINTSERVICE: checkpointing assistance -->
 <bean id="checkpointService"
   class="org.archive.crawler.framework.CheckpointService">
  </bean>

 <!-- QUEUE ASSIGNMENT POLICY -->

<!-- NAS queue assignement policy.
Note that the default H3 policy is
org.archive.crawler.frontier.SurtAuthorityQueueAssignmentPolicy
-->

 <bean id="queueAssignmentPolicy"

class="dk.netarkivet.harvester.harvesting.DomainnameQueueAssignmentPolicy">

  <property name="forceQueueAssignment" value="" /> <!-- TODO evaluate this
default -->
  <property name="deferToPrevious" value="true" />  <!-- TODO evaluate this
default -->
  <property name="parallelQueues" value="1" />      <!-- TODO evaluate this
default -->
 </bean>

 <!-- URI PRECEDENCE POLICY -->
 <!--
 <bean id="uriPrecedencePolicy"
   class="org.archive.crawler.frontier.precedence.CostUriPrecedencePolicy">
 </bean>
 -->

 <!-- COST ASSIGNMENT POLICY -->

 <bean id="costAssignmentPolicy"
   class="org.archive.crawler.frontier.UnitCostAssignmentPolicy">
 </bean>

<!-- CREDENTIAL STORE: HTTP authentication or FORM POST credentials -->
<!-- sample use of credentialStore
http://stackoverflow.com/questions/17756520/use-of-heritrixs-htmlformcrede
ntial-and-credentialstore
 -->
 <bean id="credentialStore"
   class="org.archive.modules.credential.CredentialStore">
 <property name="credentials">
 <map>
    <entry key="licitationen" value-ref="licitationen_login_1"/>
    <entry key="mymusic" value-ref="mymusic_login_1"/>
    <entry key="arto" value-ref="arto_login_1"/>
    <entry key="heerfordt" value-ref="heerfordt_login_1"/>
 </map>
</property>
```

```
</bean>

<bean id="licitationen_login_1"
class="org.archive.modules.credential.HttpAuthenticationCredential"> <!--
renamed from Rfc2617Credential -->
  <property name="domain" value="www.licitationen.dk" />
  <property name="realm" value="Dagbladet Licitationen"/>
  <property name="login" value="*****"/>
  <property name="password" value="*****"/>
</bean>

<bean id="mymusic_login_1"
class="org.archive.modules.credential.HtmlFormCredential">
  <property name="domain" value="www.mymusic.dk"/>
  <property name="loginUri"
value="http://www.mymusic.dk/konto/login2.asp"/>
  <!-- <property name="httpMethod" value="Method.POST"/> -->
  <property name="formItems">
    <map>
    <entry key="username" value="*****"/>
    <entry key="password" value="*****"/>
    <entry key="autologin" value="y"/>
    </map>
  </property>
</bean>

<bean id="arto_login_1"
class="org.archive.modules.credential.HtmlFormCredential">
  <property name="domain" value="www.arto.dk"/>
  <property name="loginUri"
value="http://www.arto.dk/r2/frames/navigation.asp"/>
  <!-- <property name="httpMethod" value="Method.POST"/> -->
  <property name="formItems">
    <map>
        <entry key="action" value="submit"/>
    <entry key="brugernavn" value="****"/>
    <entry key="kodeord" value="*****"/>
    <entry key="AutoLogin" value="ja"/>
    <entry key="loginKnap" value="Log ind"/>
    </map>
  </property>
</bean>

<bean id="heerfordt_login_1"
class="org.archive.modules.credential.HtmlFormCredential">
  <property name="domain" value="heerfordt.dk"/>
  <property name="loginUri" value="http://heerfordt.dk/"/>
<!--  <property name="http-method" value="POST"/> -->
  <property name="formItems">
    <map>
    <entry key="Brugernavn" value="*****"/>
    <entry key="Pw" value="*****"/>
    <entry key="Login" value="Login"/>
```

```
    </map>
  </property>
</bean>

<!-- sample credentials ended -->

<!-- QUOTA ENFORCER BEAN -->

<bean id="quotaenforcer"
  class="org.archive.crawler.prefetch.QuotaEnforcer">
  <property name="forceRetire" value="false"></property>

  <property name="serverMaxFetchSuccesses" value="-1"></property>
  <property name="serverMaxSuccessKb" value="-1"></property>
  <property name="serverMaxFetchResponses" value="-1"></property>
  <property name="serverMaxAllKb" value="-1"></property>

  <property name="hostMaxFetchSuccesses" value="-1"></property>
  <property name="hostMaxSuccessKb" value="-1"></property>
  <property name="hostMaxFetchResponses" value="-1"></property>
  <property name="hostMaxAllKb" value="-1"></property>

  <property name="groupMaxFetchSuccesses"
value="%{QUOTA_ENFORCER_GROUP_MAX_FETCH_SUCCES_PLACEHOLDER}">
  </property>
  <property name="groupMaxSuccessKb" value="-1"></property>
  <property name="groupMaxFetchResponses" value="-1"></property>
  <property name="groupMaxAllKb"
value="%{QUOTA_ENFORCER_MAX_BYTES_PLACEHOLDER}"></property>
 </bean>

 <!--
   REQUIRED STANDARD BEANS
    It will be very rare to replace or reconfigure the following beans.
  -->

 <!-- STATISTICSTRACKER: standard stats/reporting collector -->
 <bean id="statisticsTracker"
   class="org.archive.crawler.reporting.StatisticsTracker"
autowire="byName">
 </bean>

 <!-- CRAWLERLOGGERMODULE: shared logging facility -->
 <bean id="loggerModule"
   class="org.archive.crawler.reporting.CrawlerLoggerModule">
 </bean>

 <!-- SHEETOVERLAYMANAGER: manager of sheets of contextual overlays
      Autowired to include any SheetForSurtPrefix or
      SheetForDecideRuled beans -->
 <bean id="sheetOverlaysManager" autowire="byType"
   class="org.archive.crawler.spring.SheetOverlaysManager">
 </bean>
```

```
 <!-- BDBMODULE: shared BDB-JE disk persistence manager -->
 <bean id="bdb"
  class="org.archive.bdb.BdbModule">
 </bean>

 <!-- BDBCOOKIESTORAGE: disk-based cookie storage for FetchHTTP -->
 <bean id="cookieStorage"
   class="org.archive.modules.fetcher.BdbCookieStore">
 </bean>

 <!-- SERVERCACHE: shared cache of server/host info -->
 <bean id="serverCache"
   class="org.archive.modules.net.BdbServerCache">
 </bean>

 <!-- CONFIG PATH CONFIGURER: required helper making crawl paths relative
      to crawler-beans.cxml file, and tracking crawl files for web UI -->
 <bean id="configPathConfigurer"
   class="org.archive.spring.ConfigPathConfigurer">
 </bean>

</beans>
```