Heritrix Application Bundle

Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project (see https://webarchive.jira.com/wiki/display/Heritrix/Heritrix).

Aspire Heritrix Connector uses a custom Heritrix 3.1.1 crawl engine to crawl seed URLs based on a Heritrix job configuration file (spring application context cxml file). Instead of saving the crawled URLs to a WARC file as Heritrix would do, Aspire implements its own processor that will forward all content extracted by the crawl engine to an Aspire pipeline.

The connector, once *started*, can be *stopped*, *paused* or *resumed* sending a new Scanner Configuration Job. Typically the *start* job will contain all information required by the job to perform the scan. When pausing or stopping, the connector will wait until all the jobs it published have completed before updating the statistics and status of the connector.

Configuration

This section lists all configuration parameters available to install the Heritrix Application Bundle and to execute crawls using the connector.

General Application Configuration

Property	Type	Default	Description
snapshotDir	string	${aspire.home}/snapshots	The directory for snapshot files to be stored.
disableTextExtract	boolean	false	By default, connectors use Apache Tika to extract text from downloaded documents. If you wish to apply special text processing to the downloaded document in the workflow, you should disable text extraction. The downloaded document is then available as a content stream.
workflowReloadPeriod	int	15m	The period after which to reload the business rules. Defaults to ms, but can be suffixed with ms, s, m, h or d to indicate the required units.
workflowErrorTolerant	boolean	false	When set, exceptions in workflow rules will only effect the execution of the rule in which the exception occurs. Subsequent rules will be executed and the job will complete the workflow sucessfully. If not set, exceptions in workflow rules will be re-thrown and the job will be moved to the error workflow.
debug	Boolean	false	Controls whether debugging is enabled for the application. Debug messages will be written to the log files.

Configuration Example

To install the application bundle, add the configuration, as follows, to the *<autoStart>* section of the Aspire *settings.xml*.

```
<application config="com.searchtechnologies.aspire:app-heritrix-connector">
  <properties>
    <property name="generalConfiguration">false</property>
    <property name="heritrixJobsFolder">${app.data.dir}/heritrixJobs</property>
    <property name="jdbmDir">${app.data.dir}/incremental</property>
    <property name="checkpointIntervalMinutes">15</property>
    <property name="disableTextExtract">false</property>
    <property name="workflowReloadPeriod">15s</property>
    <property name="workflowErrorTolerant">false</property>
    <property name="debug">false</property>
  </properties>
</application>
```

Note: Any optional properties can be removed from the configuration to use the default value described on the table above.

Source Configuration

Scanner Control Configuration

The following table describes the list of attributes that the AspireObject of the incoming scanner job requires to correctly execute and control the flow of a scan process.

Element	Type	Options	Description
@action	string	start, stop, pause, resume, abort	Control command to tell the scanner which operation to perform. Use **start** option to launch a new crawl.
@actionProperties	string	full, incremental	When a **start** @action is received, it will tell the scanner to either run a **full** or an **incremental** crawl.
@normalizedCSName	string		Unique identifier name for the content source that will be crawled.
displayName	string		Display or friendly name for the content source that will be crawled.

Header Example

```
  <doc action="start" actionProperties="full" actionType="manual" crawlId="0" dbId="0" jobNumber="0"
normalizedCSName="FeedOne_Connector"
   scheduleId="0" scheduler="##AspireSystemScheduler##" sourceName="ContentSourceName">
    ...
    <displayName>testSource</displayName>
    ...
  </doc>
```

All configuration properties described in this section are relative to /doc/connectorSource of the AspireObject of the incoming Job.

Property	Type	Default	Description
defaultConfigFile	Boolean	true	Specifies the Heritrix job configuration to use for the source, **Standard Configuration** will use a default configuration file with some user specific parameters (see next properties). **Custom Configuration** will use a customized configuration file.
url	string	none	A list of seed URLs, one per line to start crawling from.
crawlScope	string	All	Selects the crawl scope for the job: **All**, **Stay within Domain** or **Stay within Host**.
maxHops	int	3	Specifies the number of allowed hops to crawl.
millisecondsPerRequest	int	3000	The number of milliseconds to wait between each request made by the Heritrix Crawl Engine during the crawl.
seedsRetry/@maxRetries	int	5	Number of retries for failed seeds.
seedsRetry/@retryDelay	int	20	Time in seconds to wait between retries for failed seeds
crawlPatterns/accept/@pattern	regex	none	Optional. A regular expression pattern to evaluate URLs against; if the URL matches the pattern, the URL is accepted by the crawler.
crawlPatterns/accept/@pattern	regex	none	Optional. A regular expression pattern to evaluate URLs against; if the URL matches the pattern, the URL is rejected by the crawler.
configFileLocation	string		Location of a custom Heritrix job configuration file (crawler-beans.cxml). This file requires the AspireHeritrixProcessor to be configured in the Disposition chain.
cleanupRegex	string		Optional. Regular Expression used to clean the content of a web page, by removing all matches of the regex in the content, before it gets to the Extract Text stage. It can be used to exclude dynamic content from index.
defaultIncrementalIndexing	boolean	false	Determines if there are custom values for incremental indexing such as *daysToDelete*, *maxFailuresToDelete*, *checkNotCrawlableContent* or *uncrawledAccessDelay*
daysToDelete	integer	2	Number of days to wait before deleting an uncrawled/not accessible URL.
maxFailuresToDelete	integer	5	Number of incremental iterations to wait before deleting an uncrawled/not accessible URL.
checkNotCrawlableContent	boolean	false	Determines if the Heritrix Scanner should verify URLs which are no longer reachable from other URLs (example: if a referring site was deleted). Otherwise those URLs will be marked as failed (and then deleted). The first time a URL is detected as not crawlable and it is still available, the scanner will send an UPDATE action for it, when it becomes crawlable again another UPDATE action will be sent for it.
uncrawledAccessDelay	integer	2000	Time in milliseconds to wait between checks (for old and failed URLs) from the same host.

fileNamePatterns /include /@pattern	regex	none	Optional. A regular expression pattern to evaluate file urls against; if the file name matches the pattern, the file is included by the scanner. Multiple include nodes can be added.
fileNamePatterns /exclude /@pattern	regex	none	Optional. A regular expression pattern to evaluate file urls against; if the file name matches the pattern, the file is excluded by the scanner. Multiple exclude nodes can be added.

Scanner Configuration Example

```
<doc action="start" actionProperties="full" normalizedCSName="ST_Web_Site">
  <connectorSource>
    <defaultConfigFile>true</defaultConfigFile>
    <url>http://www.searchtechnologies.com</url>
    <crawlScope>all</crawlScope>
    <maxHops>3</maxHops>
    <seedsRetry maxRetries="5" retryDelay="20"/>
    <millisecondsPerRequest>3000</millisecondsPerRequest>
    <crawlPatterns>
      <accept pattern="\.html$"/>
      <reject pattern="\.js$"/>
      <reject pattern="\.css$"/>
    </crawlPatterns>
    <defaultIncrementalIndexing>true</defaultIncrementalIndexing>
    <fetchDelay>500</fetchDelay>
    <daysToDelete>2</daysToDelete>
    <maxFailuresToDelete>5</maxFailuresToDelete>
    <checkNotCrawlableContent>true</checkNotCrawlableContent>
    <uncrawledAccessDelay>2000</uncrawledAccessDelay>
    <cleanupRegex><!--googleoff: all-->[\s\S]*<!--googleon: all--></cleanupRegex>
    <fileNamePatterns>
      <include pattern=".*"/>
      <exclude pattern=".*robots.txt.*"/>
    </fileNamePatterns>
  </connectorSource>
</doc>
```

or using a custom crawler beans:

```
<doc action="start" actionProperties="full" normalizedCSName="ST_Web_Site">
  <connectorSource>
    <defaultConfigFile>false</defaultConfigFile>
    <configFileLocation>config/custom-crawler-beans.cxml</configFileLocation>
    <defaultIncrementalIndexing>true</defaultIncrementalIndexing>
    <fetchDelay>500</fetchDelay>
    <daysToDelete>2</daysToDelete>
    <maxFailuresToDelete>5</maxFailuresToDelete>
    <checkNotCrawlableContent>true</checkNotCrawlableContent>
    <uncrawledAccessDelay>2000</uncrawledAccessDelay>
    <cleanupRegex><!--googleoff: all-->[\s\S]*<!--googleon: all--></cleanupRegex>
    <fileNamePatterns>
      <include pattern=".*"/>
      <exclude pattern=".*robots.txt.*"/>
    </fileNamePatterns>
  </connectorSource>
</doc>
```

Note: To launch a crawl, the job should be sent (processed/enqueued) to the "/HeritrixConnector/Main" pipeline.

Output

```
<doc>
  <docType>item</docType>
  <url>http://www.searchtechnologies.com/</url>
  <id>http://www.searchtechnologies.com/</id>
  <fetchUrl>http://www.searchtechnologies.com/</fetchUrl>
  <displayUrl>http://www.searchtechnologies.com/</displayUrl>
  <snapshotUrl>001 http://www.searchtechnologies.com/</snapshotUrl>
  <sourceType>heritrix</sourceType>
  <sourceName>ST_Web_Site</sourceName>
  <connectorSpecific type="heritrix">
    <field name="md5">IFAUKQSBGFCUCNJVGMYTOQ2FHA4EEMRWIJDDCMJWII4TQM2CGIZA</field>
    <field name="xslt">false</field>
    <field name="discoveredBy"/>
    <field name="pathFromSeed"/>
  </connectorSpecific>
  <connectorSource>
    <defaultConfigFile>true</defaultConfigFile>
    <url>http://www.searchtechnologies.com</url>
    <crawlScope>all</crawlScope>
    <maxHops>3</maxHops>
    <seedsRetry maxRetries="5" retryDelay="20"/>
    <millisecondsPerRequest>3000</millisecondsPerRequest>
    <crawlPatterns/>
    <defaultIncrementalIndexing>false</defaultIncrementalIndexing>
    <cleanupRegex><!--googleoff: all-->[\s\S]*<!--googleon: all--></cleanupRegex>
    <fileNamePatterns>
      <include pattern=".*"/>
      <exclude pattern=".*robots.txt.*"/>
    </fileNamePatterns>
    <displayName>ST Web Site</displayName>
  </connectorSource>
  <action>add</action>
  <hierarchy>
    <item id="ADDFF324E6D09222031F87DA77854D50" level="1" name="ST_Web_Site" url="http://www.searchtechnologies.
com/"/>
  </hierarchy>
  <title source="ExtractTextStage/title">The Enterprise Search Implementation Experts</title>
  <contentType source="ExtractTextStage/Content-Type">application/xhtml+xml</contentType>
  <description source="ExtractTextStage/description">Search Technologies is the largest IT services company
dedicated to
    enterprise search implementation, consulting, and managed services. Our expertise covers all leading search
products,
    and all aspects of search applications.
  </description>
  <keywords source="ExtractTextStage/keywords">Enterprise Search, Search Engine Experts, Consulting</keywords>
  <extension source="ExtractTextStage">
    <field name="Content-Location">http://www.searchtechnologies.com/</field>
    <field name="Content-Encoding">ISO-8859-1</field>
    <field name="resourceName">http://www.searchtechnologies.com/</field>
    <field name="google-site-verification">jP1bIfjuuyZUYfTkYc_O6ZlTHxCm07voTDcMk72Z8oQ</field>
    <field name="dc:title">The Enterprise Search Implementation Experts</field>
  </extension>
  <content source="ExtractTextStage">
   The Enterprise Search Experts
     .
     .
     .
  </content>
</doc>
```

Heritrix Configuration File

Standard Configuration

- Sets the Seed URLs to the TextSeedModule bean.
- Uses the following Decide Rules to configure the crawl scope (in this order):
    - **RejectDecideRule**

        - REJECT
    - **SurtPrefixedDecideRule**
        - ACCEPT
    - **MatchesListRegexDecideRule**
        - ACCEPT all URLs that match a regex in the list of accept patterns configured by the user.
    - **FetchStatusMatchesRegexDecideRule**
        - ACCEPT all URLs with fetch status of 200-300
    - **FetchStatusMatchesRegexDecideRule**
        - REJECT all URLs with fetch status of 400-500
    - **TooManyHopsDecideRule**
        - REJECT all URLs after the number of maximum hops defined by the user.
    - **TransclusionDecideRule**
        - ACCEPT
    - **NotOnDomainsDecideRule/NotOnHostsDecideRule**
        - Depending on users choice:
            - All (NotOnDomainsDecideRule -> ACCEPT)
            - Stay within Domain (NotOnDomainsDecideRule -> REJECT)
            - Stay within Host (NotOnHostsDecideRule -> REJECT)
    - **SurtPrefixedDecideRule**
        - REJECT those configured on the negative-surts.dump file, initially empty
    - **MatchesListRegexDecideRule**
        - REJECT all URLs that match a regex in the list of reject patterns configured by the user.
    - **PathologicalPathDecideRule**
        - REJECT
    - **TooManyPathSegmentsDecideRule**
        - REJECT
    - **PrerequisiteAcceptDecideRule**
        - ACCEPT (robots.txt for example)
    - **SchemeNotInSetDecideRule**
        - REJECT
- Uses the AspireHeritrixProcessor on the disposition chain

```
 <bean id="aspireProcessor" class="com.searchtechnologies.aspire.components.heritrixconnector.
AspireHeritrixProcessor"/>
```

Custom Configuration

The custom configuration file can be configured to use any Heritrix feature available for the standalone version, but instead of using the WARCWriterProcessor as the first step on the DispositionChain, it requires the AspireHeritrixProcessor:

```
<bean id="aspireProcessor" class="com.searchtechnologies.aspire.components.heritrixconnector.
AspireHeritrixProcessor"/>
```

It is required to configure the following digest properties, since they are used for incremental indexing.

```
<bean id="fetchHttp" class="org.archive.modules.fetcher.FetchHTTP">
 <property name="digestContent" value="true" />
 <property name="digestAlgorithm" value="md5" />
</bean>
```

Example configuration file for Aspire Heritrix Connector: Crawler-beans.xml

Aspire Heritrix Connector uses a custom Heritrix Engine that can handle NTLM authentication. For details on how to configure NTLM authentication see Using a Custom Heritrix Configuration File

Our custom Heritrix engine can handle (if desired) XSL transformations to extract links from XSLT generated HTML, and will also perform the data extraction from the HTML, not the original XML. For details on how to enable XSLT see Using a Custom Heritrix Configuration File