Heritrix FAQ & Troubleshooting

After a few minutes, there are no updates submitted in the content source statistics

- This is normally a DNS issue. Check with your network team to see if the site URL(s) you want to crawl are reachable from the server running Aspire.

Where can I learn more about Heritrix Architecture and configuring?

- You can go to http://crawler.archive.org/Mohr-et-al-2004.pdf paper to have a better overview of the Heritrix architecture.

Where to look at if my Heritrix Connector is not working as expected?

One of the most common cause of confusion when using the Aspire Heritrix Connector is to detect where the issues should be fixed: is it an Aspire connector issue? or is it a Heritrix Crawl engine issue?.

You can detect which part is the problem by looking at what do each part does:

- The Heritrix Connector is only encharged of:
  - Preparing the crawler-beans.cxml file with the user's parameters.
  - Create a Heritrix Engine Job and start the crawl.
  - Receive the crawled Web Pages or documents from the Heritrix Crawl Engine and send them to an Aspire Pipeline.
  - Manage the incremental indexing (ignore unchanged documents, send new documents and delete the ones that are no longer accessible) from the documents received from the Heritrix Crawl Engine.
  - Cleanup content of each document using the *Cleanup Regex*.
  - Apply the Index include/exclude patterns to the documents received.

- The Heritrix Crawl engine is encharged of:
  - Actually perform the crawl.
  - Check for robots policies.
  - Perform authentication (including NTLM in our custom engine) if specified and required.
  - XSLT transformation (in out custom engine).
  - Fetch an input stream for each document.
  - Calculate an MD5 digest of the content used later by the Aspire Heritrix Connector to do the incremental indexing.
  - Apply the Crawl patterns specified by the user.

My crawl is very slow

Consider that Heritrix always try to protect the servers it is crawling, by throttling the requests to the same hostname with the maxDelays.

Try with a lower *millisecondsPerRequest* if you are sending the configuration via an Aspire Job, or the following property:

```
<bean class="org.archive.crawler.postprocessor.DispositionProcessor" id="disposition">
  <property name="maxDelayMs" value="3000"/>
</bean>
```

You can also increase the number of parallel connections to the same hostname. See more information at Using a Custom Heritrix Configuration File at the Configuring Concurrent Connections to the same hostname section

More crawler beans configuration at: https://webarchive.jira.com/wiki/display/Heritrix/Basic+Crawl+Job+Settings .

Why is does an incremental crawl last as long as a full crawl?

The Heritrix Connector performs incremental crawls based on a disk-backed HashMap, which have the exact documents that have been indexed by the connector to the search engine associated with a content digest signature. On an incremental crawl the connector fully crawls the web sites configured the same way as a full crawl, but it only indexes the modified, new or deleted documents during that crawl.

Why am I only getting 6000 documents discovered per URL?

Heritrix by default sets a maximum of 6000 links to extract from a single URL, the rest of the links found are discarded an therefore not crawled. You can configure that by changing a bean inside a custom heritrix crawler beans. See more information on how to configure that at Using a Custom Heritrix Configuration File

More information about Heritrix Connector

For more information on configuring the crawler beans for custom features see Using a Custom Heritrix Configuration File.

For general FAQ of Heritrix go to http://crawler.archive.org/faq.html .

If you are interested in developing new features in Heritrix go tohttp://crawler.archive.org/articles/developer_manual/index.html .

To request new features or more information please contact us at http://www.searchtechnologies.com/contacts.html .