Project Goals

- Setup a web-crawler with Heritrix
- Archive files from vt.edu
- Integrate with Wayback
- Set-up Search with Solr (Stretch)
Problems Encountered

● Older version of software.
● Finding documentation to configure Heritrix.
  ○ Only crawl vt.edu pages.
  ○ Crawl all vt.edu pages.
● Issues with CentOS firewalls.
Work Accomplished

- Working set-up of Heritrix that successfully crawls vt.edu web-pages.
  - Customized configuration to increase crawl depth.
  - Reject non-domain based URLs.

- Working set-up of Wayback machine:
  - Processes warc files from Heritrix.
  - Front-end for Heritrix-based crawls.
Lessons Learned

● Sometimes, documentation leaves much to be desired.

● Crawls can be extremely large if not configured properly.
Demo

Heritrix:
- https://administrator:mQW8GzEsZAr8SxAketPY@webarchive.cc.vt.edu:12222/

Wayback:
- http://webarchive.cc.vt.edu/
Engine

Memory:
162426 KiB used, 232960 KiB current heap, 232960 KiB max heap

Jobs Directory:
/apps/nas/heritrix-3.2.0/jobs

Job Directories

(7) detected rescan

vteduDevRun1 «Finished: ABORTED» 2 launches
/apps/nas/heritrix-3.2.0/jobs/vteduDevRun1/crawler-beans.cxml
(last at 2014-05-05T23:35:39.836Z)

vteduDevRun «Finished: ABORTED» 1 launches
/apps/nas/heritrix-3.2.0/jobs/vteduDevRun/crawler-beans.cxml
(last at 2014-05-05T22:44:36.438Z)

vtedurun «Finished: FINISHED» 1 launches
/apps/nas/heritrix-3.2.0/jobs/vtedurun/crawler-beans.cxml
(last at 2014-04-28T18:44:38.186Z)

vtedu-2 0 launches
/apps/nas/heritrix-3.2.0/jobs/vtedu-2/profile-crawler-beans.cxml

1 launches
Job vteduDevRun1

(2 launches, last 3d3h ago)

Job Log more

2014-05-07T01:11:32.815Z INFO FINISHED 20140505233541
2014-05-07T01:11:24.301Z WARNING unable to tally host stats for 0.0.0.0 (in thread 'org.archive.crawler.frontier...
2014-05-07T01:10:31.907Z INFO STOPPING 20140505233541
2014-05-07T01:10:27.275Z INFO PAUSING 20140505233541
2014-05-06T15:59:52.912Z WARNING unable to tally host stats for http://0.0.0.0/most-content-tag.js (in thread 'T

Job is Finished: ABORTED

Totals
2,520,583 downloaded + 7,325,359 queued = 9,845,942 total
547 GiB crawled (547 GiB novel, 0 B dupByHash, 764 B notModified)

Alerts
18 tail alert log...

Rates
0 URLs/sec (27.35 avg), 0 KB/sec (8,228 avg)

Load
Job is Finished: ABORTED

Totals
  2,520,583 downloaded + 7,325,359 queued = 9,845,942 total
  547 GiB crawled (547 GiB novel, 0 B dupByHash, 764 B notModified)

Alerts
  18 tail alert log...

Rates
  0 URLs/sec (27.35 avg); 0 KB/sec (8,228 avg)

Load
  0 active of 0 threads; 99.36 congestion ratio; 167,453 deepest queue; 88 average depth

Elapsed
  1d1h35m49s15ms

Threads
  n/a

Frontier
  FINISH - 90,029 URI queues: 903 active (0 in-process; 69 ready; 834 snoozed); 81,963 inactive; 0 ineligible; 0 retired; 7,163 exhausted

Memory
  167277 KiB used; 232960 KiB current heap; 232960 KiB max heap

Crawl Log more
<beans xmlns="http://www.springframework.org/schema/beans"
      xmlns:context="http://www.springframework.org/schema/context"
      xmlns:tx="http://www.springframework.org/schema/tx"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

</context:annotation-config/>

<bean id="simpleOverride" class="org.springframework.beans.factory.config.PropertyOverrideConfigurer">
  <property name="properties">
    <value>
      # This Properties map is specified in the Java 'property list' text format
      # http://java.sun.com/j2se/1.4/docs/api/java/util/Properties.html#load%28java.io.Reader%29
      
      metadata.operatorContactUrl=http://www.vt.edu
      metadata.foName=basio
      metadata.description=Basic crawl starting with useful defaults
      
      ## more entries ##
    </value>
  </property>
</bean>

<bean id="longerOverride" class="org.springframework.beans.factory.config.PropertyOverrideConfigurer">
  <property name="properties">
    <props>
      <prop key="source.textSource.value">

      </prop>
    </props>
  </property>
</bean>

<bean id="override3" class="org.springframework.beans.factory.config.PropertyOverrideConfigurer">
  <property name="properties">
    <props>
      <prop key="source.textSource.value">

      </prop>
    </props>
  </property>
</bean>

<bean id="override2" class="org.springframework.beans.factory.config.PropertyOverrideConfigurer">
  <property name="properties">
    <props>
      <prop key="source.textSource.value">

      </prop>
    </props>
  </property>
</bean>

<bean id="override1" class="org.springframework.beans.factory.config.PropertyOverrideConfigurer">
  <property name="properties">
    <props>
      <prop key="source.textSource.value">

      </prop>
    </props>
  </property>
</bean>

</beans>
This is the new Wayback Machine prototype. Any URL in ARC files accessible to this service can be searched above.

Home | Help
Spotlight on INNOVATION

The battle against brain cancer

A neurologist at the veterinary college is researching new treatments for aggressive brain tumors in dogs and humans.
Questions?