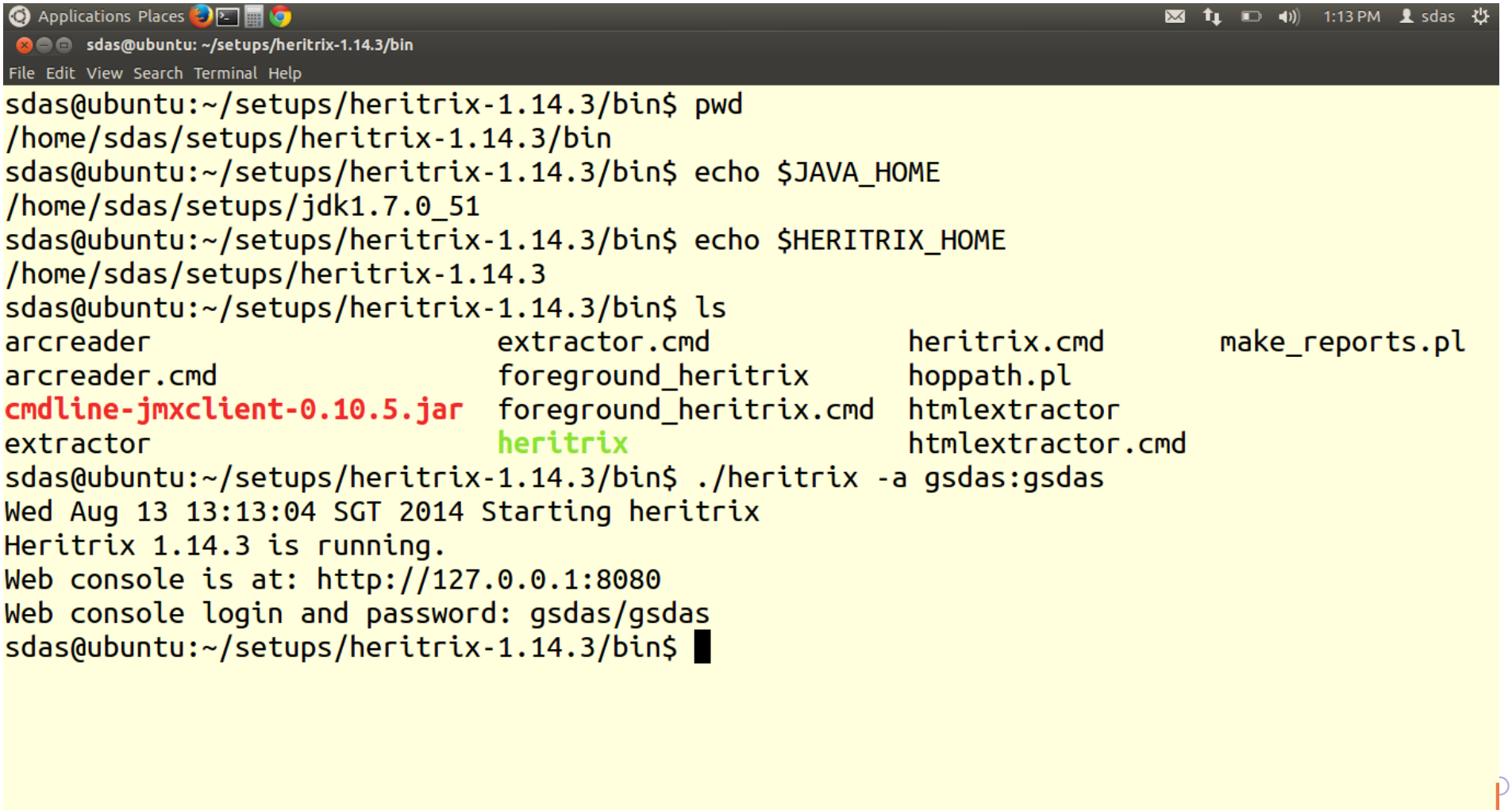
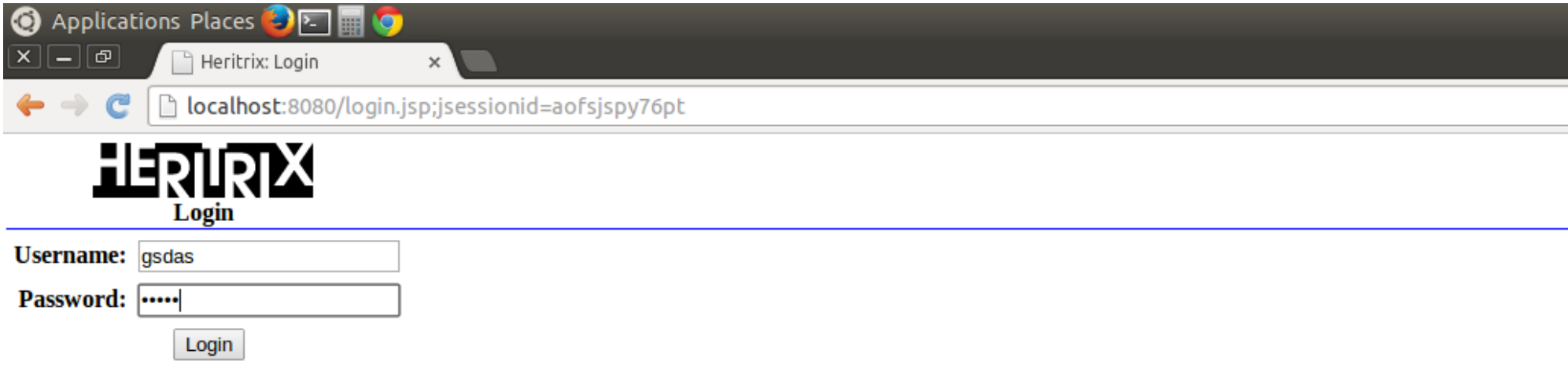
Why Heritrix?

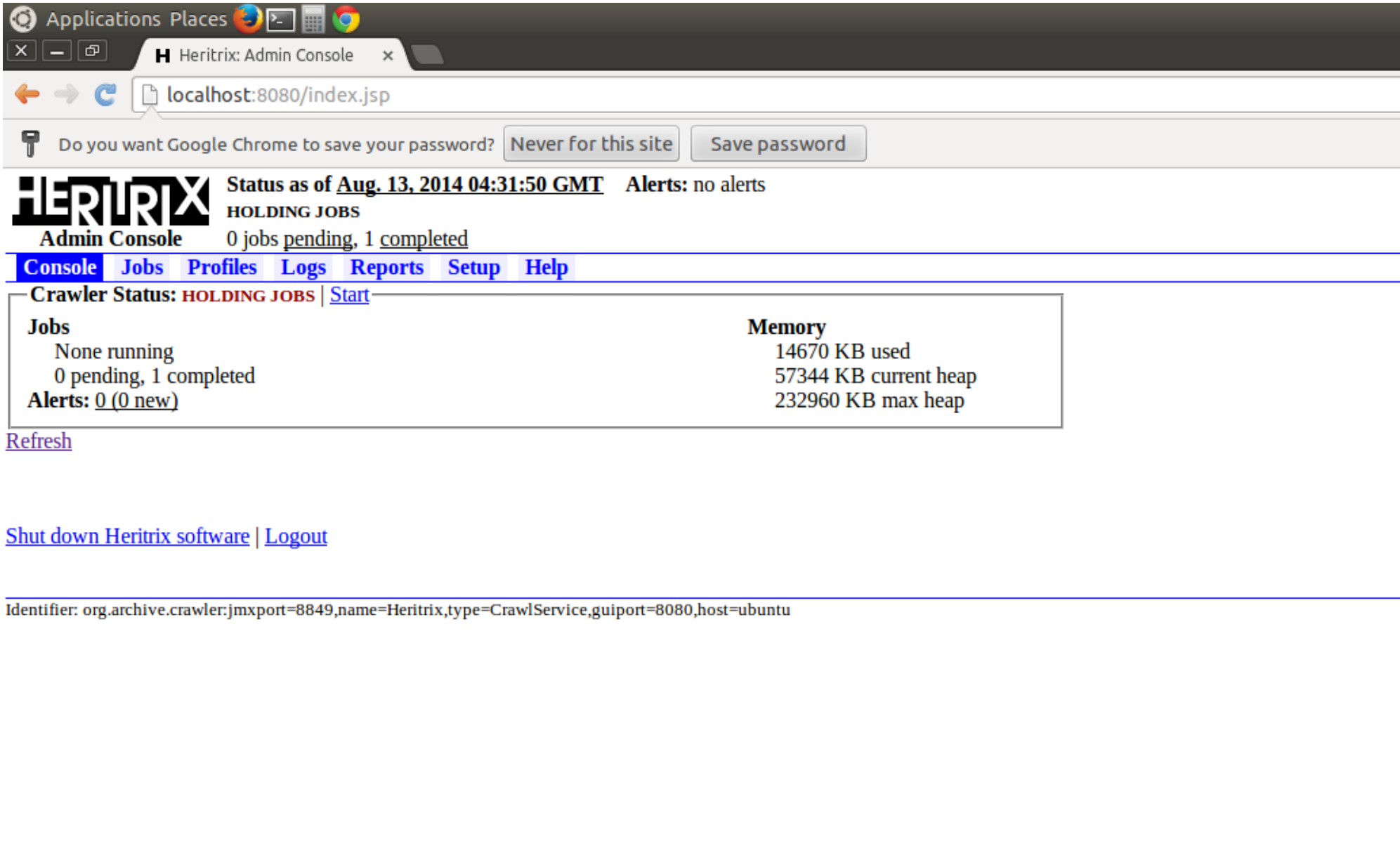
- Internet Archive's web-scale, archival-quality web crawler project
- Open-source and extensible
- Written in Java and used in CiteSeer

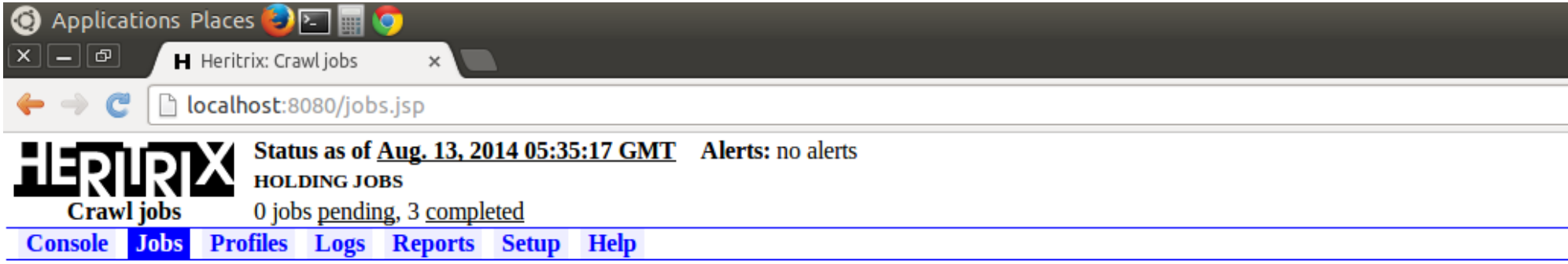
Download/untar/cd bin

http://crawler.archive.org/index.html Go to sourceforge downloads page and get version 1.14.3




Applications Places
Heritrix: Login
localhost:8080/login.jsp;jsessionid=aofsjspy76pt
HERITRIX
Login
Username:
gsdas
Password:
Login


Applications Places
Heritrix: Admin Console
localhost:8080/index.jsp
Do you want Google Chrome to save your password?
Never for this site
Save password
HERITRIX
Admin Console
Status as of Aug. 13, 2014 04:31:50 GMT
Alerts: no alerts
HOLDING JOBS
0 jobs pending, 1 completed
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Crawler Status: HOLDING JOBS | Start
Jobs
None running
0 pending, 1 completed
Alerts: 0 (0 new)
Memory
14670 KB used
57344 KB current heap
232960 KB max heap
Refresh
Shut down Heritrix software | Logout
Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=8080,host=ubuntu



HERITRIX

Crawl jobs

Status as of Aug. 13, 2014 05:35:17 GMT **Alerts:** no alerts

HOLDING JOBS

0 jobs pending, 3 completed

Console | Jobs | Profiles | Logs | Reports | Setup | Help

Create New Job

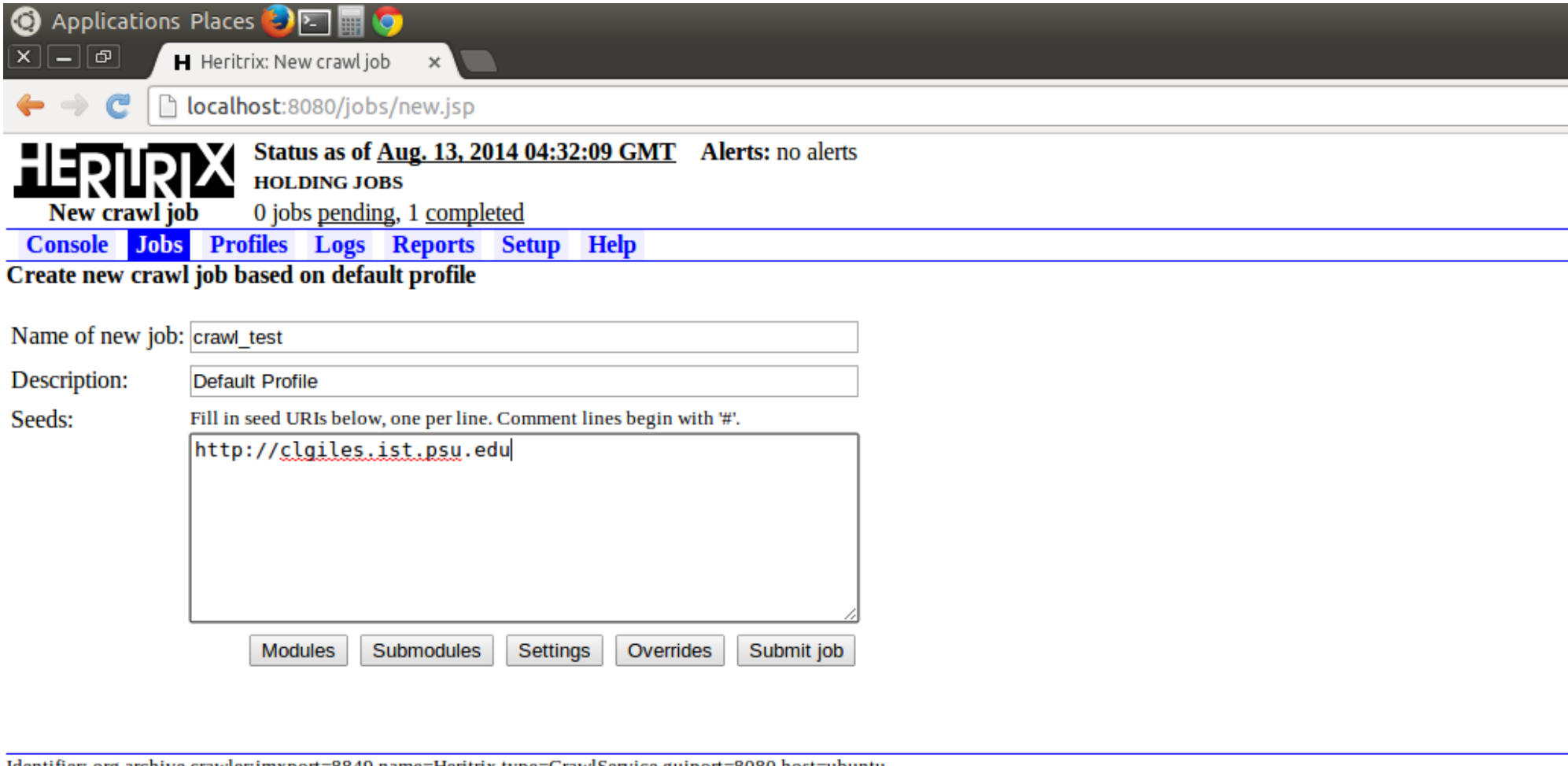
- Based on existing job
- Based on a recovery
- Based on a profile
- With defaults

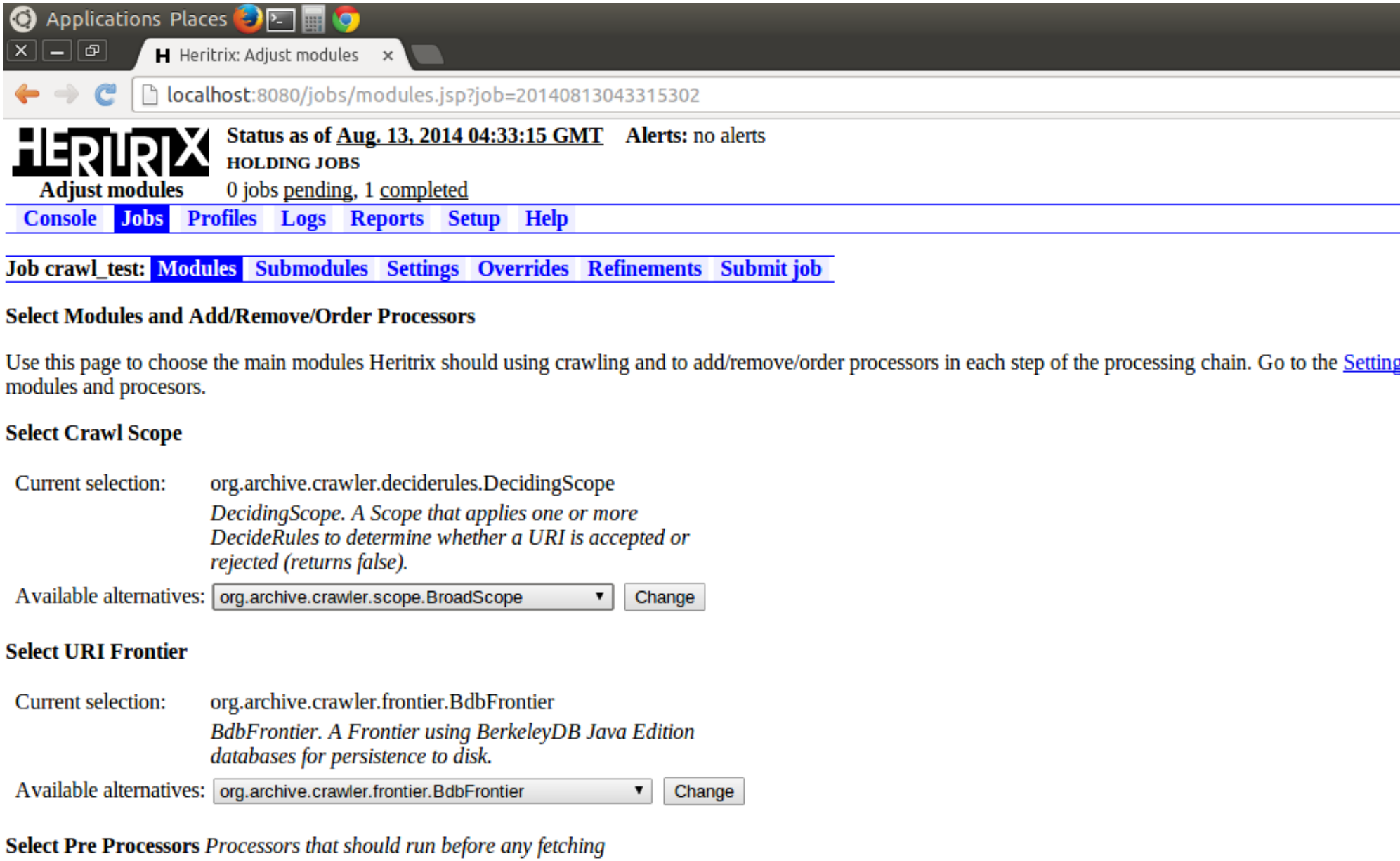
Pending Jobs (0)

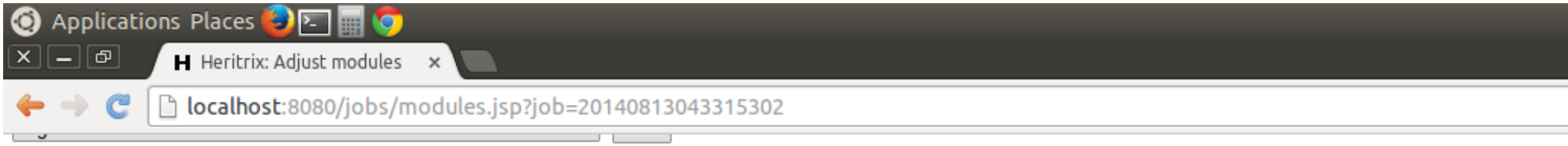
Completed Jobs(3)

UID	Name	Status	Options						
20140813044606002	mirror_crawl	*Finished - Ended by operator*	Crawl order	Crawl report	Seeds report	Seed file	Logs	Journal	Delete
20140813043422142	crawl_test	*Finished - Ended by operator*	Crawl order	Crawl report	Seeds report	Seed file	Logs	Journal	Delete
20140812043421791	default	*Finished - Ended by operator*	Crawl order	Crawl report	Seeds report	Seed file	Logs	Journal	Delete

Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=8080,host=ubuntu


Applications Places
Heritrix: New crawl job
localhost:8080/jobs/new.jsp
HERITRIX
New crawl job
Status as of Aug. 13, 2014 04:32:09 GMT
Alerts: no alerts
HOLDING JOBS
0 jobs pending, 1 completed
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Create new crawl job based on default profile
Name of new job:
crawl_test
Description:
Default Profile
Seeds:
Fill in seed URIs below, one per line. Comment lines begin with '#'.
http://clgiles.ist.psu.edu
Modules
Submodules
Settings
Overrides
Submit job
Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=8080,host=ubuntu


Applications Places
Heritrix: Adjust modules
localhost:8080/jobs/modules.jsp?job=20140813043315302
HERITRIX
Adjust modules
Status as of Aug. 13, 2014 04:33:15 GMT
Alerts: no alerts
HOLDING JOBS
0 jobs pending, 1 completed
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Job crawl_test:
Modules
Submodules
Settings
Overrides
Refinements
Submit job
Select Modules and Add/Remove/Order Processors
Use this page to choose the main modules Heritrix should using crawling and to add/remove/order processors in each step of the processing chain. Go to the Setting
modules and procesors.
Select Crawl Scope
Current selection:
org.archive.crawler.deciderules.DecidingScope
DecidingScope. A Scope that applies one or more DecideRules to determine whether a URI is accepted or rejected (returns false).
Available alternatives:
org.archive.crawler.scope.BroadScope
Change
Select URI Frontier
Current selection:
org.archive.crawler.frontier.BdbFrontier
BdbFrontier. A Frontier using BerkeleyDB Java Edition databases for persistence to disk.
Available alternatives:
org.archive.crawler.frontier.BdbFrontier
Change
Select Pre Processors Processors that should run before any fetching



Select Fetchers *Processors that fetch documents using various protocols*

org.archive.crawler.fetcher.FetchDNS		Down	Remove Info
org.archive.crawler.fetcher.FetchHTTP	Up		Remove Info

org.archive.crawler.prefetch.Preselector ▾ Add

Select Extractors *Processors that extracts links from URIs*

org.archive.crawler.extractor.ExtractorHTTP		Down	Remove Info
org.archive.crawler.extractor.ExtractorHTML	Up	Down	Remove Info
org.archive.crawler.extractor.ExtractorCSS	Up	Down	Remove Info
org.archive.crawler.extractor.ExtractorJS	Up	Down	Remove Info
org.archive.crawler.extractor.ExtractorSWF	Up		Remove Info

org.archive.crawler.prefetch.Preselector ▾ Add

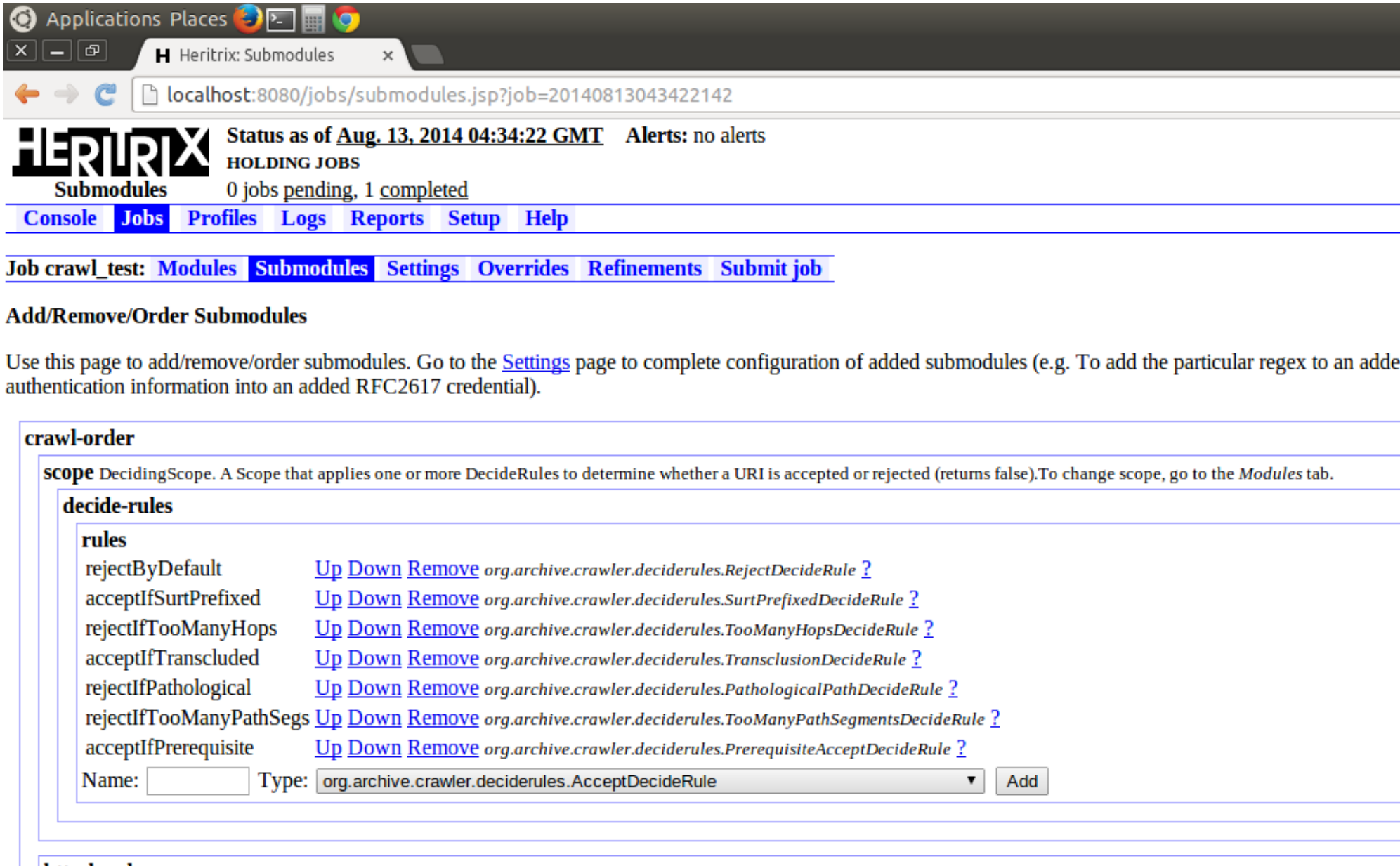
Select Writers *Processors that write documents to archive files*

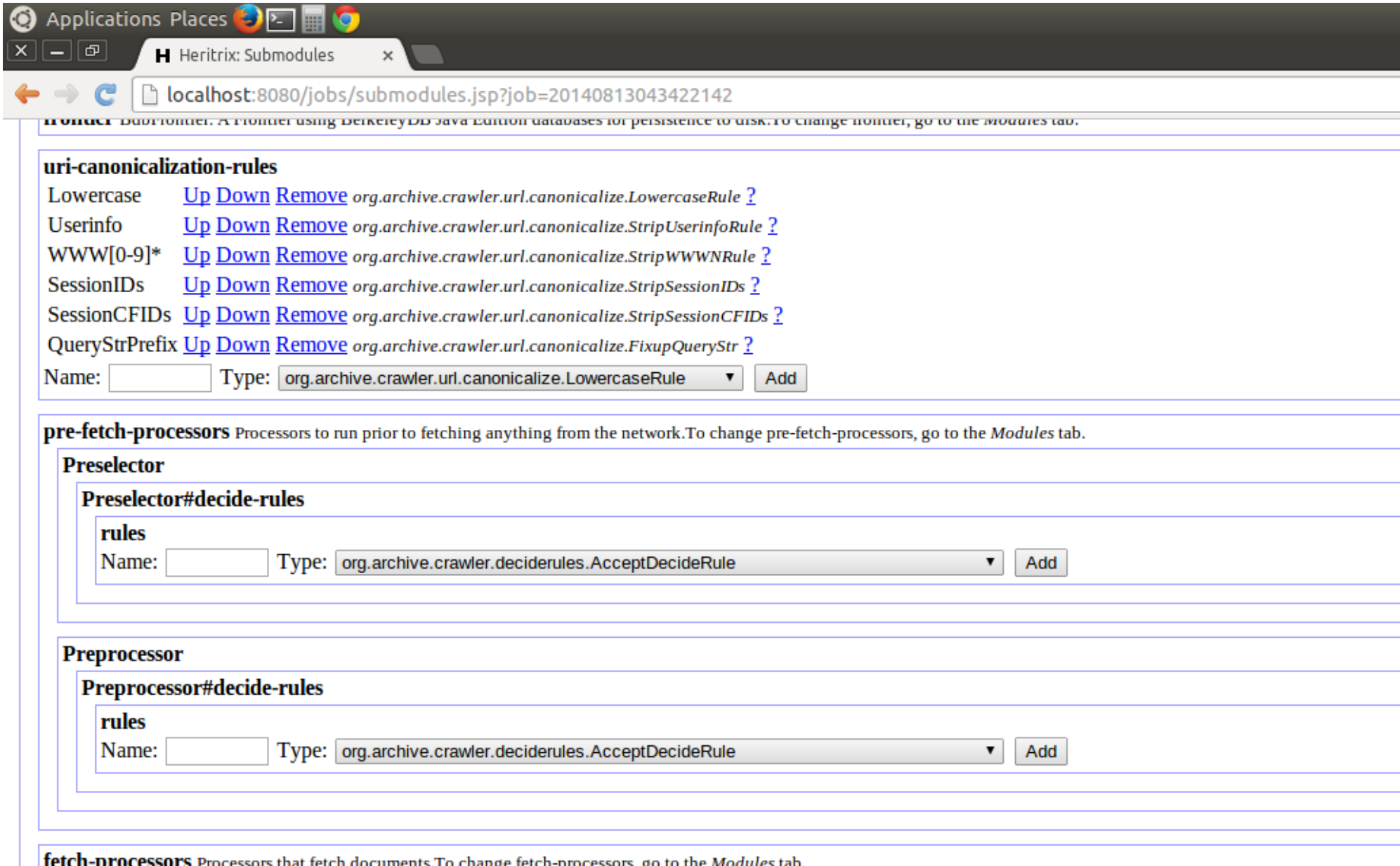
org.archive.crawler.writer.ARCWriterProcessor			Remove Info

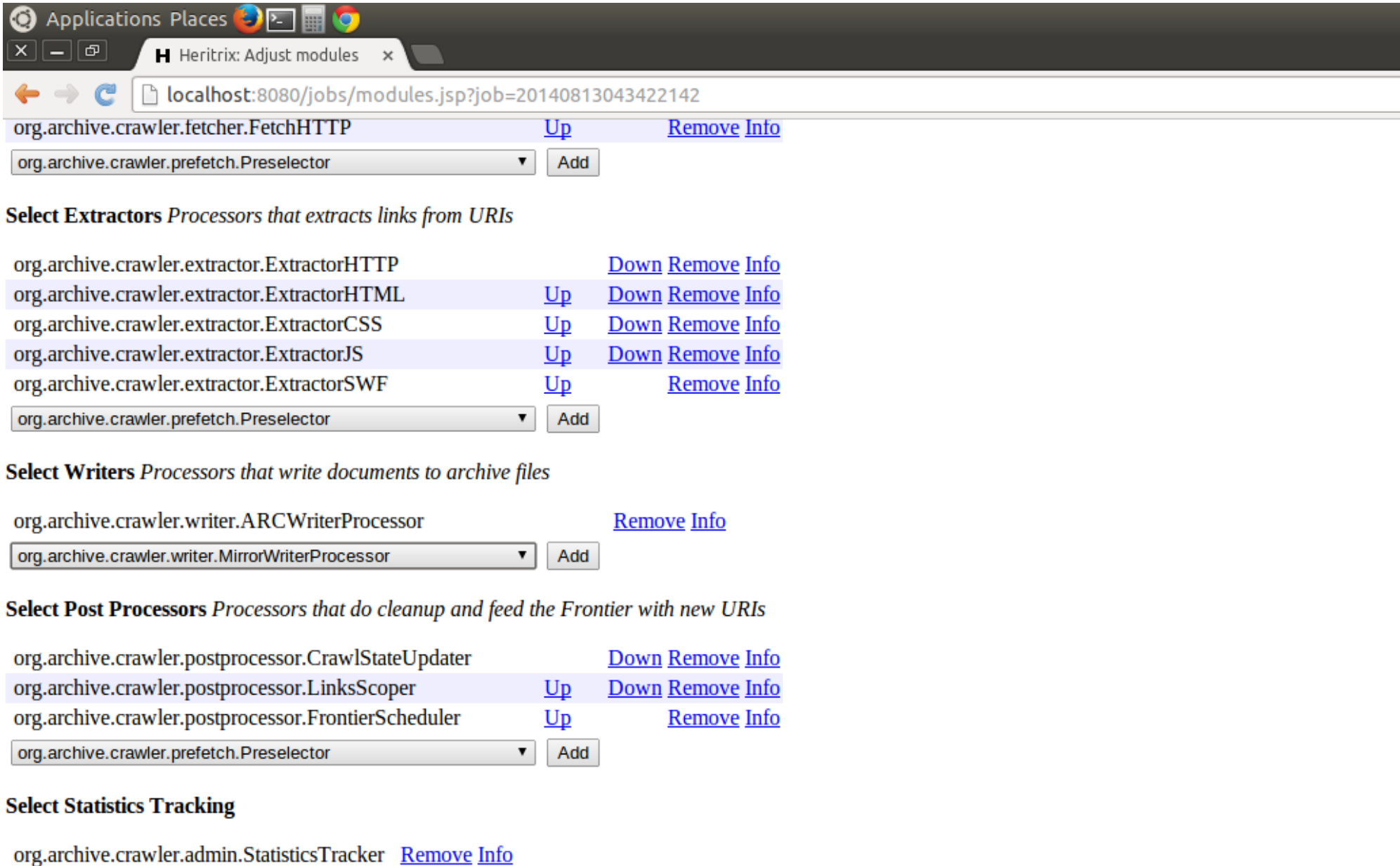
org.archive.crawler.prefetch.Preselector ▾ Add

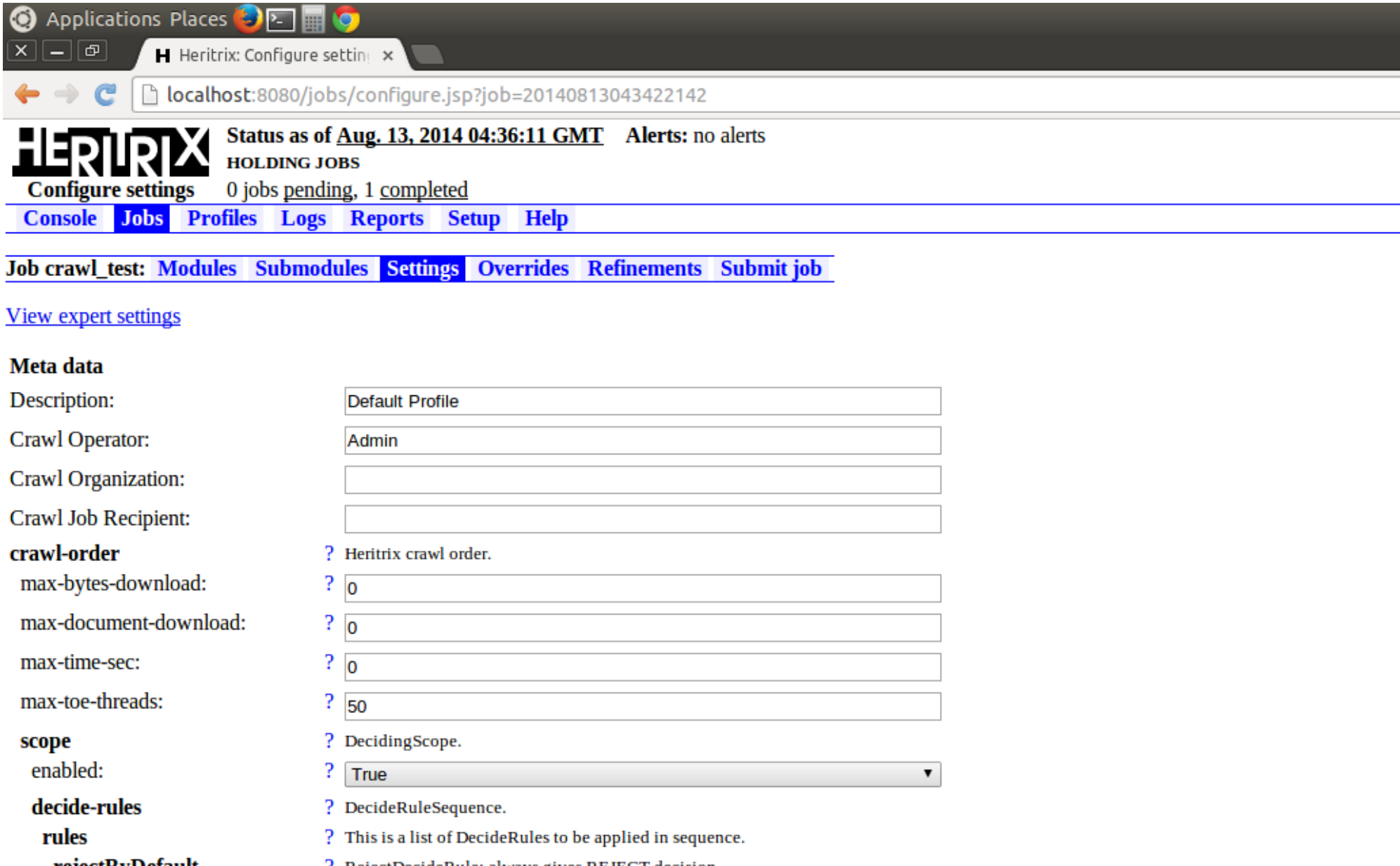
Select Post Processors *Processors that do cleanup and feed the Frontier with new URIs*

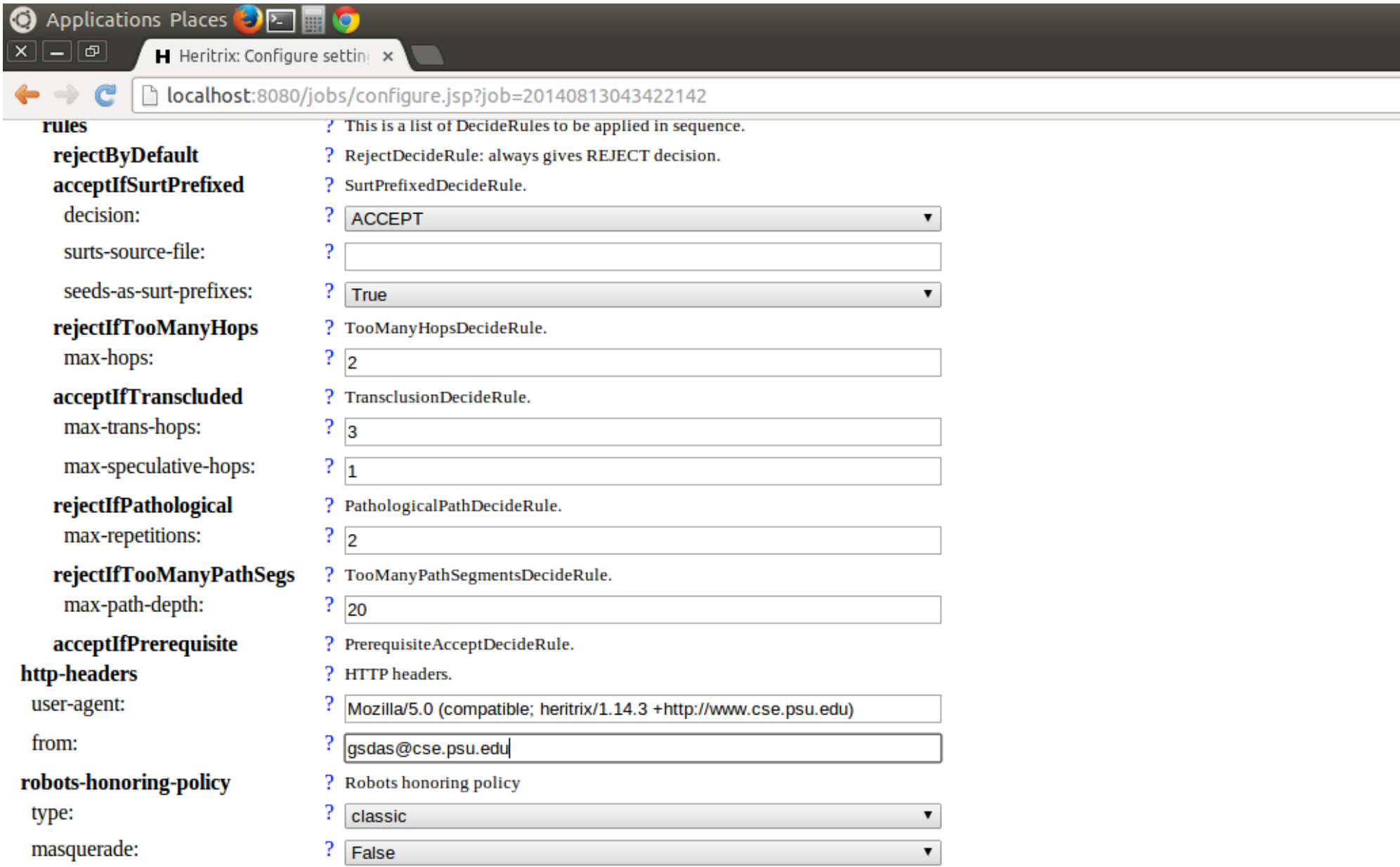
org.archive.crawler.postprocessor.CrawlStateUpdater		Down	Remove Info
org.archive.crawler.postprocessor.LinksScoper	Up	Down	Remove Info
org.archive.crawler.postprocessor.FrontierScheduler	Up		Remove Info

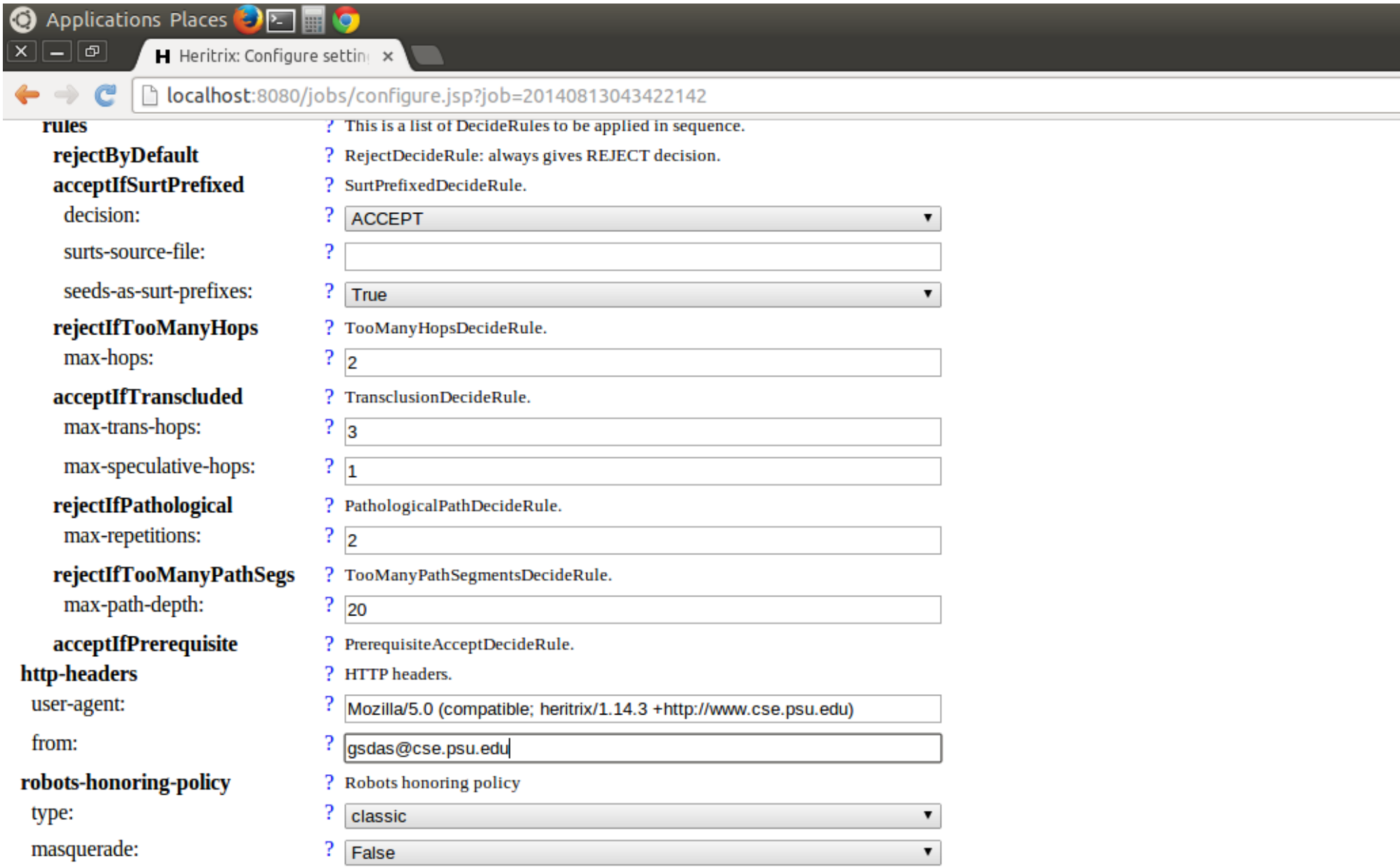

Applications Places
Heritrix: Submodules
localhost:8080/jobs/submodules.jsp?job=20140813043422142
HERITRIX
Submodules
Status as of Aug. 13, 2014 04:34:22 GMT
Alerts: no alerts
HOLDING JOBS
0 jobs pending, 1 completed
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Job crawl_test:
Modules
Submodules
Settings
Overrides
Refinements
Submit job
Add/Remove/Order Submodules
Use this page to add/remove/order submodules. Go to the Settings page to complete configuration of added submodules (e.g. To add the particular regex to an adde
authentication information into an added RFC2617 credential).
crawl-order
scope DecidingScope. A Scope that applies one or more DecideRules to determine whether a URI is accepted or rejected (returns false).To change scope, go to the Modules tab.
decide-rules
rules
rejectByDefault Up Down Remove org.archive.crawler.deciderules.RejectDecideRule ?
acceptIfSurtPrefixed Up Down Remove org.archive.crawler.deciderules.SurtPrefixedDecideRule ?
rejectIfTooManyHops Up Down Remove org.archive.crawler.deciderules.TooManyHopsDecideRule ?
acceptIfTranscluded Up Down Remove org.archive.crawler.deciderules.TransclusionDecideRule ?
rejectIfPathological Up Down Remove org.archive.crawler.deciderules.PathologicalPathDecideRule ?
rejectIfTooManyPathSegs Up Down Remove org.archive.crawler.deciderules.TooManyPathSegmentsDecideRule ?
acceptIfPrerequisite Up Down Remove org.archive.crawler.deciderules.PrerequisiteAcceptDecideRule ?
Name:
Type: org.archive.crawler.deciderules.AcceptDecideRule
Add

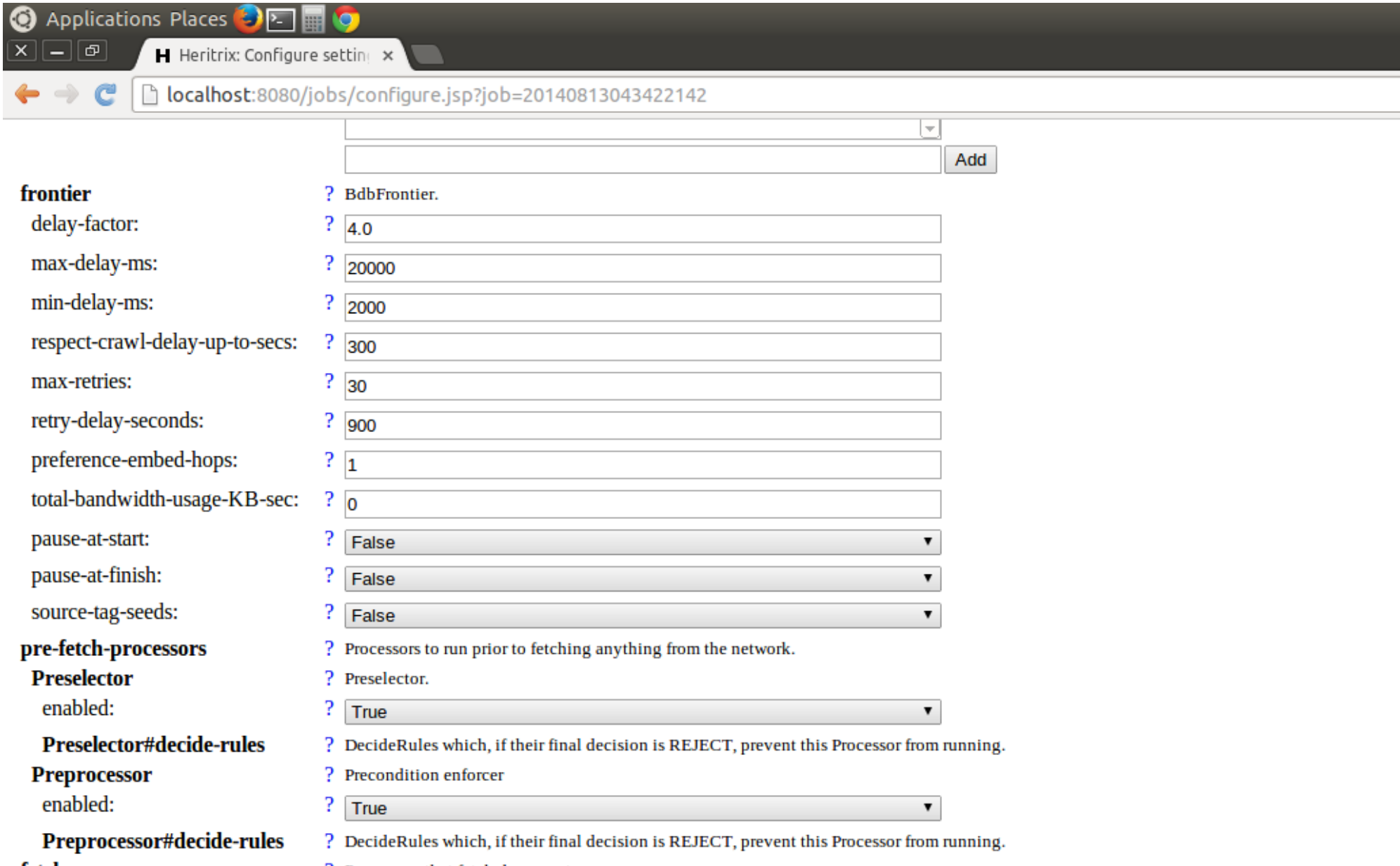

Applications Places
Heritrix: Submodules
localhost:8080/jobs/submodules.jsp?job=20140813043422142
uri-canonicalization-rules
Lowercase Up Down Remove org.archive.crawler.url.canonicalize.LowercaseRule ?
Userinfo Up Down Remove org.archive.crawler.url.canonicalize.StripUserinfoRule ?
WWW[0-9]* Up Down Remove org.archive.crawler.url.canonicalize.StripWWWNRule ?
SessionIDs Up Down Remove org.archive.crawler.url.canonicalize.StripSessionIDs ?
SessionCFIDs Up Down Remove org.archive.crawler.url.canonicalize.StripSessionCFIDs ?
QueryStrPrefix Up Down Remove org.archive.crawler.url.canonicalize.FixupQueryStr ?
Name: Type: org.archive.crawler.url.canonicalize.LowercaseRule Add
pre-fetch-processors Processors to run prior to fetching anything from the network.To change pre-fetch-processors, go to the Modules tab.
Preselector
Preselector#decide-rules
rules
Name: Type: org.archive.crawler.deciderules.AcceptDecideRule Add
Preprocessor
Preprocessor#decide-rules
rules
Name: Type: org.archive.crawler.deciderules.AcceptDecideRule Add
fetch-processors Processors that fetch documents To change fetch-processors, go to the Modules tab

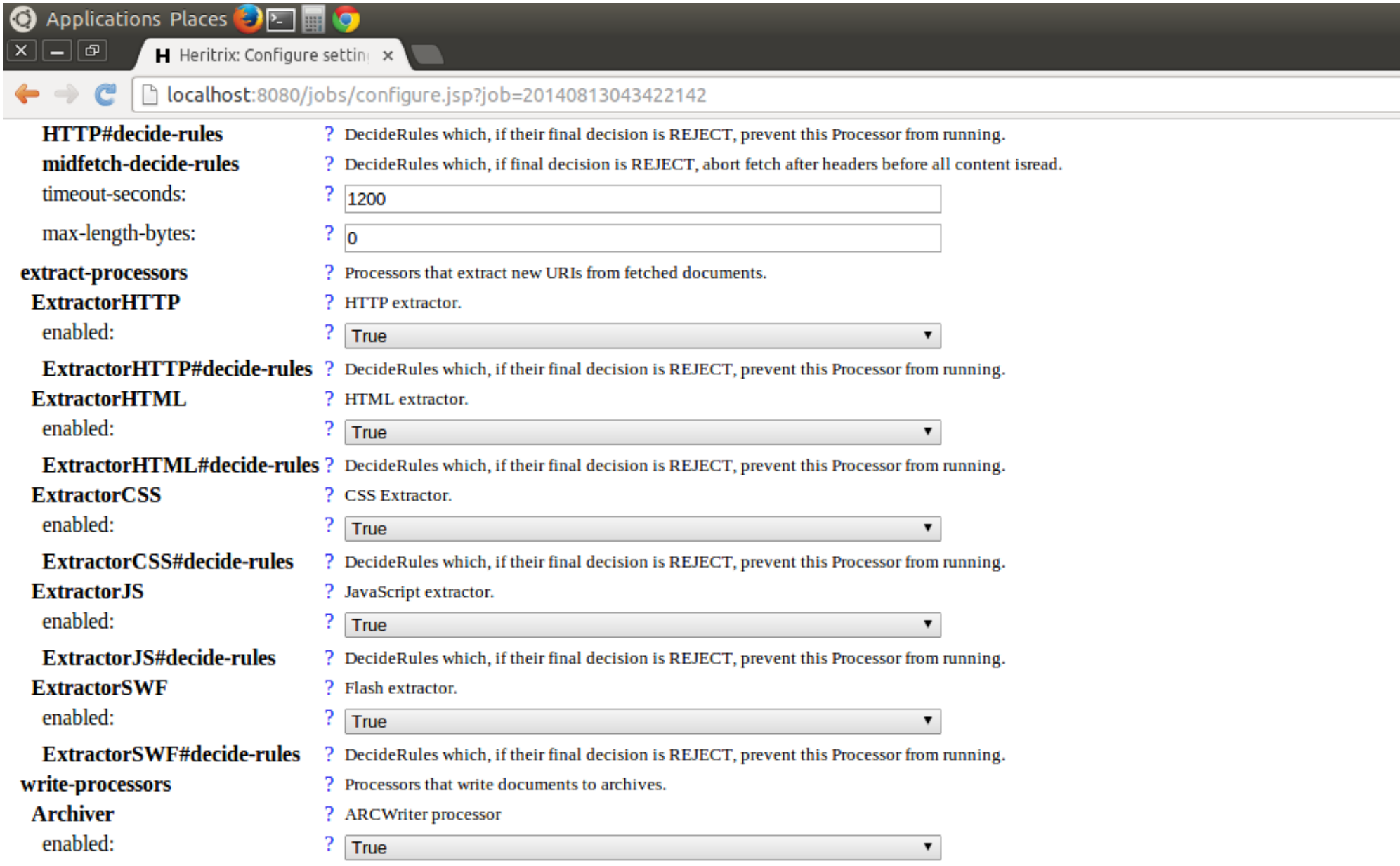

Applications Places
Heritrix: Adjust modules
localhost:8080/jobs/modules.jsp?job=20140813043422142
org.archive.crawler.fetcher.FetchHTTP Up Remove Info
org.archive.crawler.prefetch.Preselector Add
Select Extractors Processors that extracts links from URIs
org.archive.crawler.extractor.ExtractorHTTP Down Remove Info
org.archive.crawler.extractor.ExtractorHTML Up Down Remove Info
org.archive.crawler.extractor.ExtractorCSS Up Down Remove Info
org.archive.crawler.extractor.ExtractorJS Up Down Remove Info
org.archive.crawler.extractor.ExtractorSWF Up Remove Info
org.archive.crawler.prefetch.Preselector Add
Select Writers Processors that write documents to archive files
org.archive.crawler.writer.ARCWriterProcessor Remove Info
org.archive.crawler.writer.MirrorWriterProcessor Add
Select Post Processors Processors that do cleanup and feed the Frontier with new URIs
org.archive.crawler.postprocessor.CrawlStateUpdater Down Remove Info
org.archive.crawler.postprocessor.LinksScoper Up Down Remove Info
org.archive.crawler.postprocessor.FrontierScheduler Up Remove Info
org.archive.crawler.prefetch.Preselector Add
Select Statistics Tracking
org.archive.crawler.admin.StatisticsTracker Remove Info

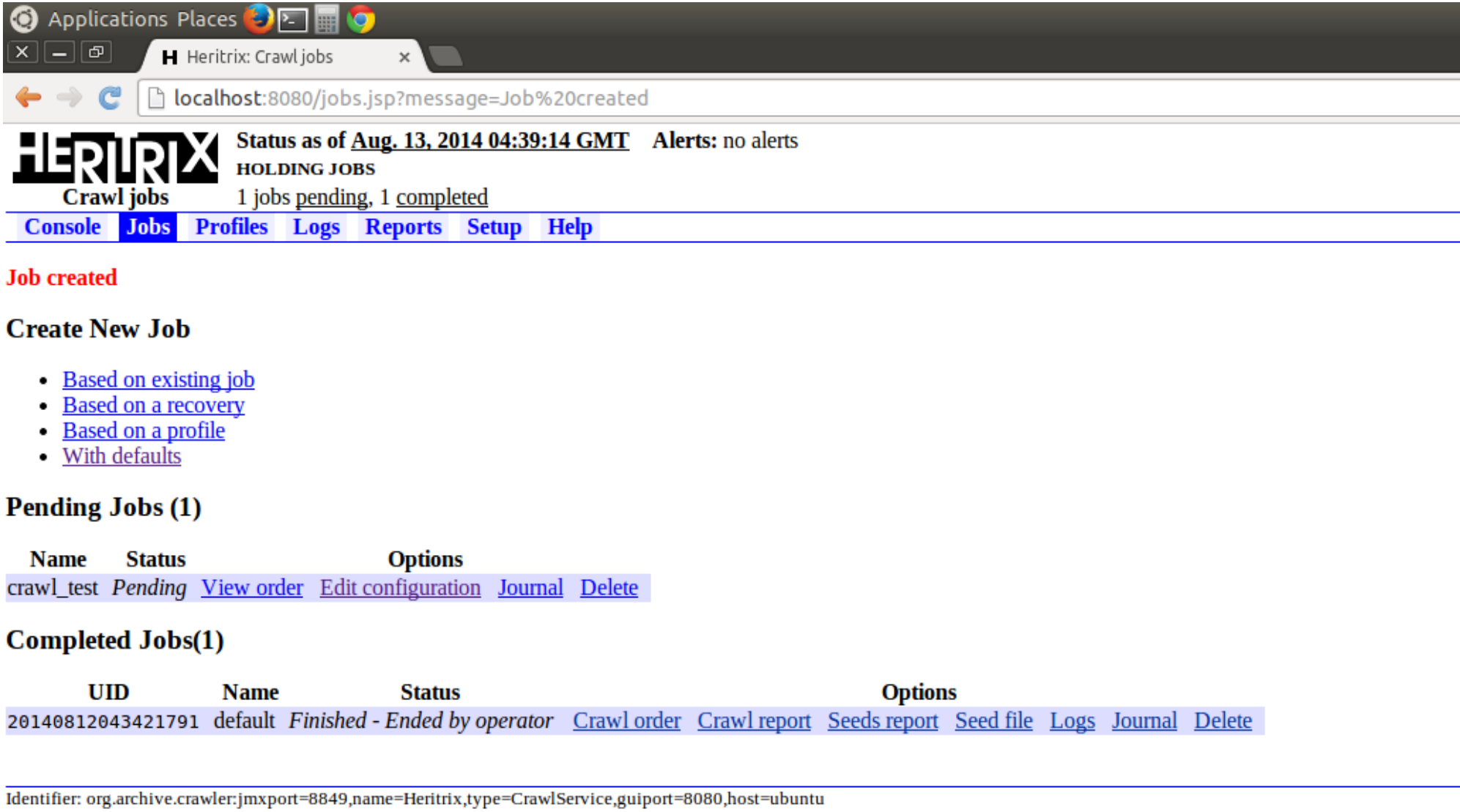

Applications Places
Heritrix: Configure settin
localhost:8080/jobs/configure.jsp?job=20140813043422142
HERITRIX
Configure settings
Status as of Aug. 13, 2014 04:36:11 GMT
Alerts: no alerts
HOLDING JOBS
0 jobs pending, 1 completed
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Job crawl_test:
Modules
Submodules
Settings
Overrides
Refinements
Submit job
View expert settings
Meta data
Description:
Default Profile
Crawl Operator:
Admin
Crawl Organization:
Crawl Job Recipient:
crawl-order
Heritrix crawl order.
max-bytes-download:
0
max-document-download:
0
max-time-sec:
0
max-toe-threads:
50
scope
DecidingScope.
enabled:
True
decide-rules
DecideRuleSequence.
rules
This is a list of DecideRules to be applied in sequence.

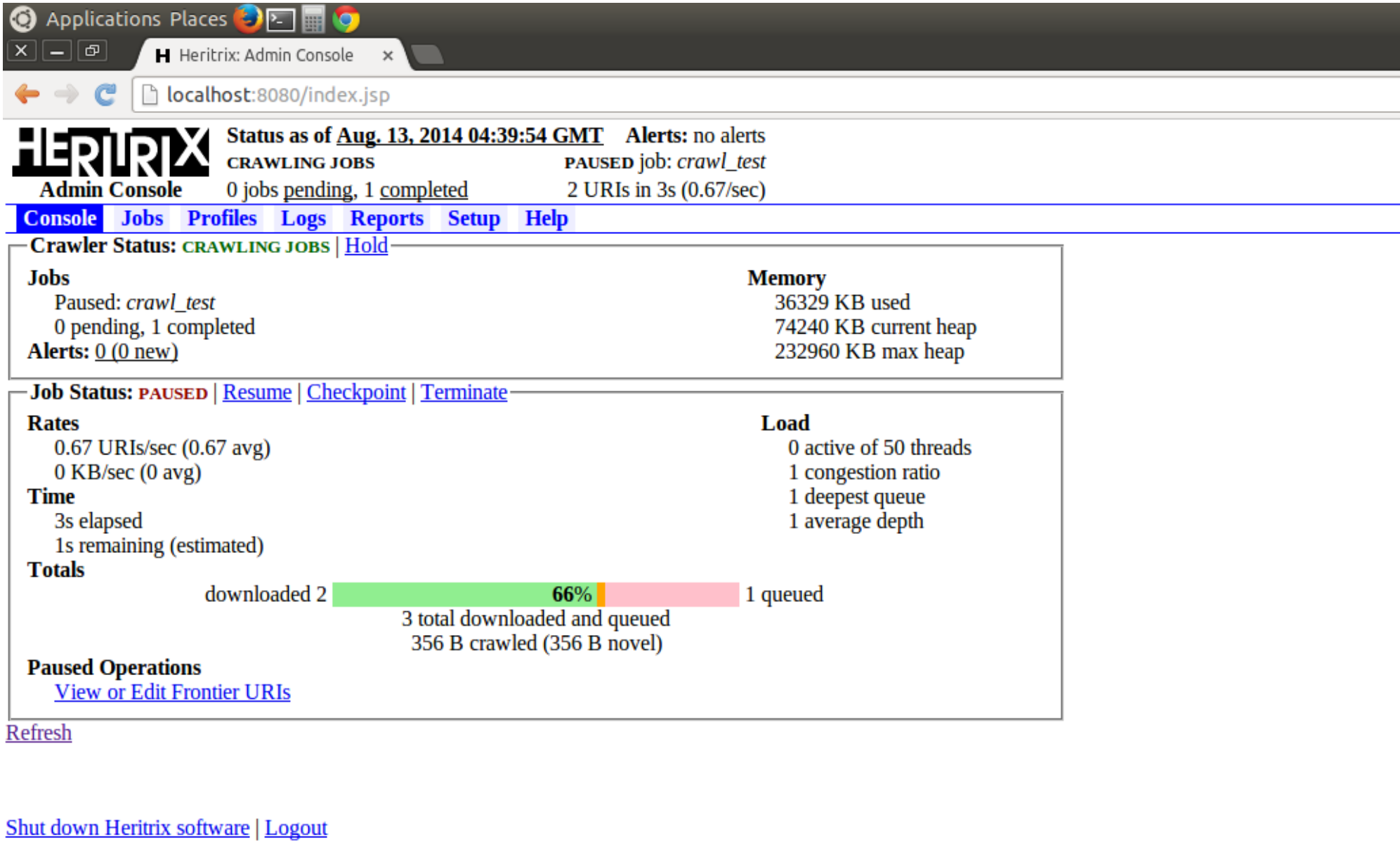

Applications Places
Heritrix: Configure settin
localhost:8080/jobs/configure.jsp?job=20140813043422142
rules ? This is a list of DecideRules to be applied in sequence.
rejectByDefault ? RejectDecideRule: always gives REJECT decision.
acceptIfSurtPrefixed ? SurtPrefixedDecideRule.
decision: ? ACCEPT
surts-source-file: ?
seeds-as-surt-prefixes: ? True
rejectIfTooManyHops ? TooManyHopsDecideRule.
max-hops: ? 2
acceptIfTranscluded ? TransclusionDecideRule.
max-trans-hops: ? 3
max-speculative-hops: ? 1
rejectIfPathological ? PathologicalPathDecideRule.
max-repetitions: ? 2
rejectIfTooManyPathSegs ? TooManyPathSegmentsDecideRule.
max-path-depth: ? 20
acceptIfPrerequisite ? PrerequisiteAcceptDecideRule.
http-headers ? HTTP headers.
user-agent: ? Mozilla/5.0 (compatible; heritrix/1.14.3 +http://www.cse.psu.edu)
from: ? gsdas@cse.psu.edu
robots-honoring-policy ? Robots honoring policy
type: ? classic
masquerade: ? False

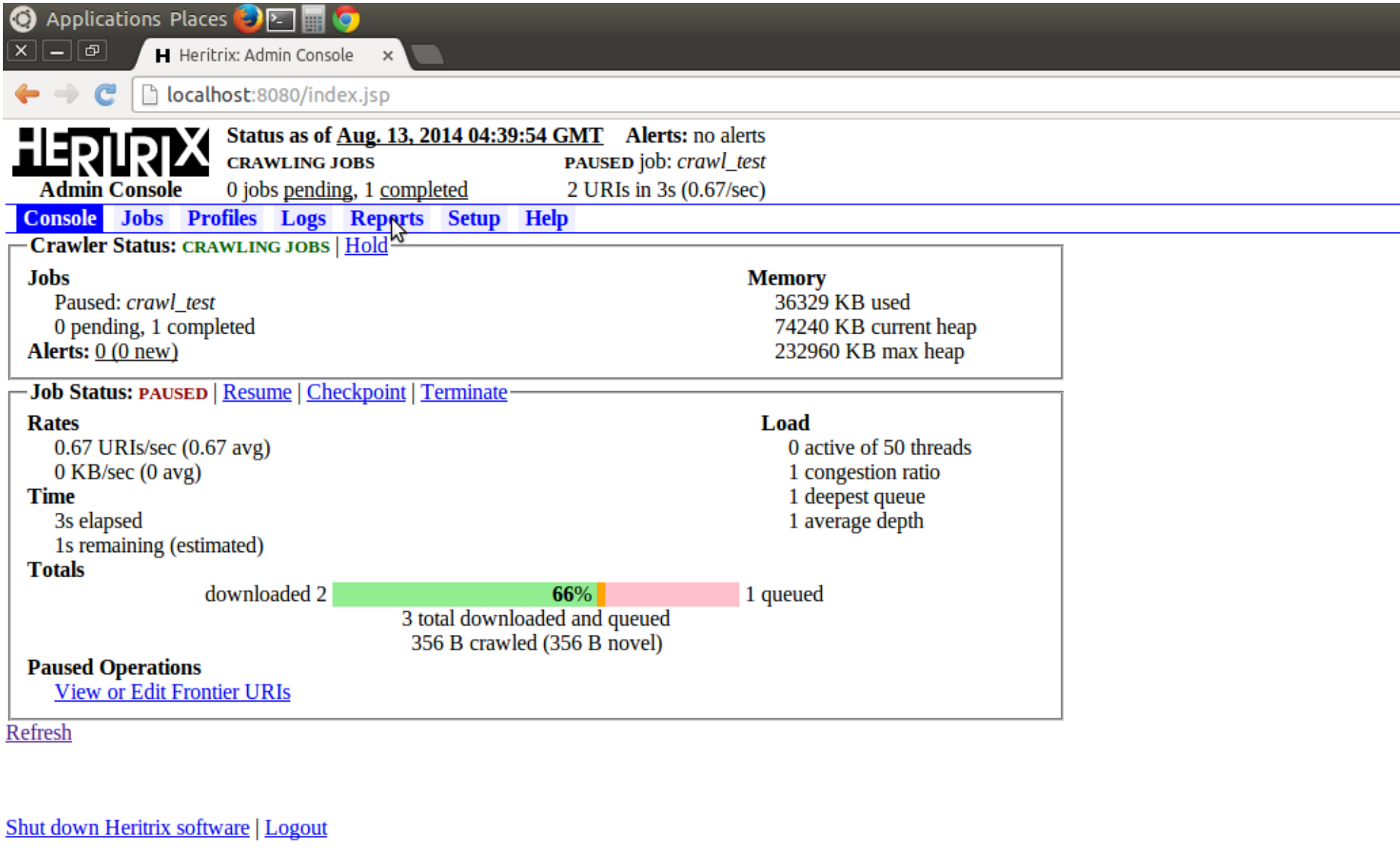

Applications Places
Heritrix: Configure settin
localhost:8080/jobs/configure.jsp?job=20140813043422142
rules ? This is a list of DecideRules to be applied in sequence.
rejectByDefault ? RejectDecideRule: always gives REJECT decision.
acceptIfSurtPrefixed ? SurtPrefixedDecideRule.
decision: ? ACCEPT
surts-source-file: ?
seeds-as-surt-prefixes: ? True
rejectIfTooManyHops ? TooManyHopsDecideRule.
max-hops: ? 2
acceptIfTranscluded ? TransclusionDecideRule.
max-trans-hops: ? 3
max-speculative-hops: ? 1
rejectIfPathological ? PathologicalPathDecideRule.
max-repetitions: ? 2
rejectIfTooManyPathSegs ? TooManyPathSegmentsDecideRule.
max-path-depth: ? 20
acceptIfPrerequisite ? PrerequisiteAcceptDecideRule.
http-headers ? HTTP headers.
user-agent: ? Mozilla/5.0 (compatible; heritrix/1.14.3 +http://www.cse.psu.edu)
from: ? gsdas@cse.psu.edu
robots-honoring-policy ? Robots honoring policy
type: ? classic
masquerade: ? False

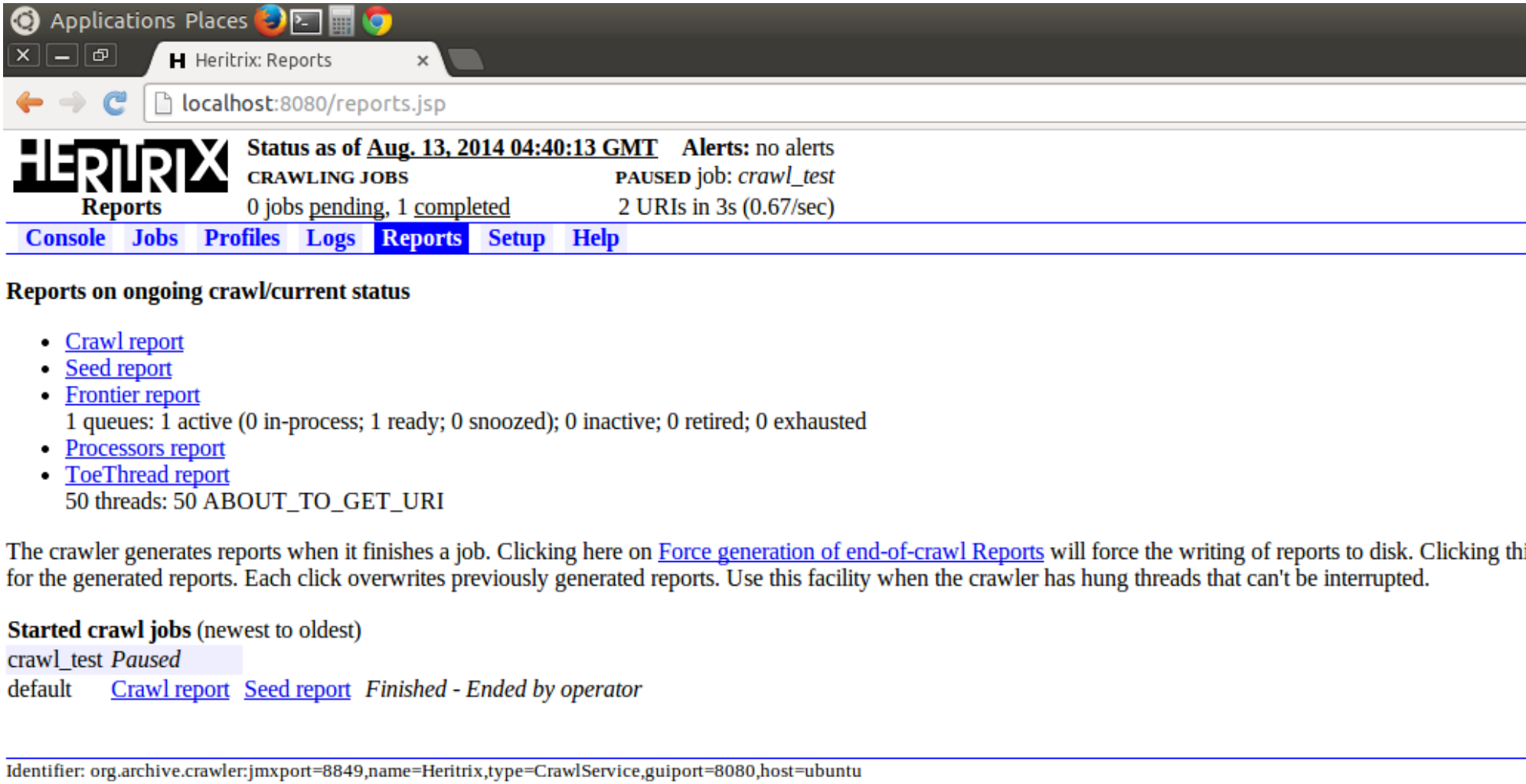

Applications Places
Heritrix: Configure settin
localhost:8080/jobs/configure.jsp?job=20140813043422142
Add
frontier ? BdbFrontier.
delay-factor: ? 4.0
max-delay-ms: ? 20000
min-delay-ms: ? 2000
respect-crawl-delay-up-to-secs: ? 300
max-retries: ? 30
retry-delay-seconds: ? 900
preference-embed-hops: ? 1
total-bandwidth-usage-KB-sec: ? 0
pause-at-start: ? False
pause-at-finish: ? False
source-tag-seeds: ? False
pre-fetch-processors ? Processors to run prior to fetching anything from the network.
Preselector ? Preselector.
enabled: ? True
Preselector#decide-rules ? DecideRules which, if their final decision is REJECT, prevent this Processor from running.
Preprocessor ? Precondition enforcer
enabled: ? True
Preprocessor#decide-rules ? DecideRules which, if their final decision is REJECT, prevent this Processor from running.

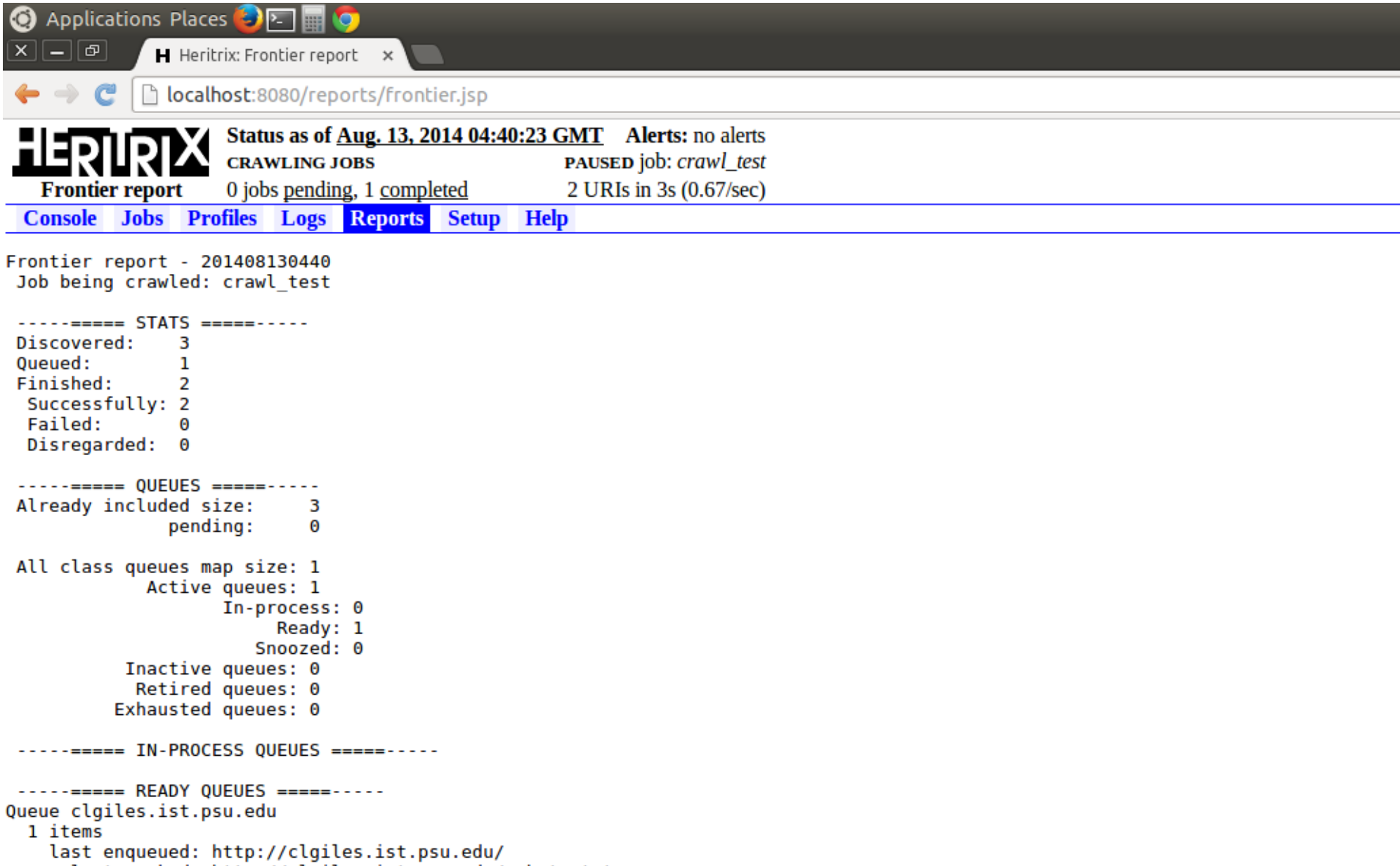

Applications Places
Heritrix: Configure settin
localhost:8080/jobs/configure.jsp?job=20140813043422142
HTTP#decide-rules ? DecideRules which, if their final decision is REJECT, prevent this Processor from running.
midfetch-decide-rules ? DecideRules which, if final decision is REJECT, abort fetch after headers before all content isread.
timeout-seconds: ? 1200
max-length-bytes: ? 0
extract-processors ? Processors that extract new URIs from fetched documents.
ExtractorHTTP ? HTTP extractor.
enabled: ? True
ExtractorHTTP#decide-rules ? DecideRules which, if their final decision is REJECT, prevent this Processor from running.
ExtractorHTML ? HTML extractor.
enabled: ? True
ExtractorHTML#decide-rules ? DecideRules which, if their final decision is REJECT, prevent this Processor from running.
ExtractorCSS ? CSS Extractor.
enabled: ? True
ExtractorCSS#decide-rules ? DecideRules which, if their final decision is REJECT, prevent this Processor from running.
ExtractorJS ? JavaScript extractor.
enabled: ? True
ExtractorJS#decide-rules ? DecideRules which, if their final decision is REJECT, prevent this Processor from running.
ExtractorSWF ? Flash extractor.
enabled: ? True
ExtractorSWF#decide-rules ? DecideRules which, if their final decision is REJECT, prevent this Processor from running.
write-processors ? Processors that write documents to archives.
Archiver ? ARCWriter processor
enabled: ? True

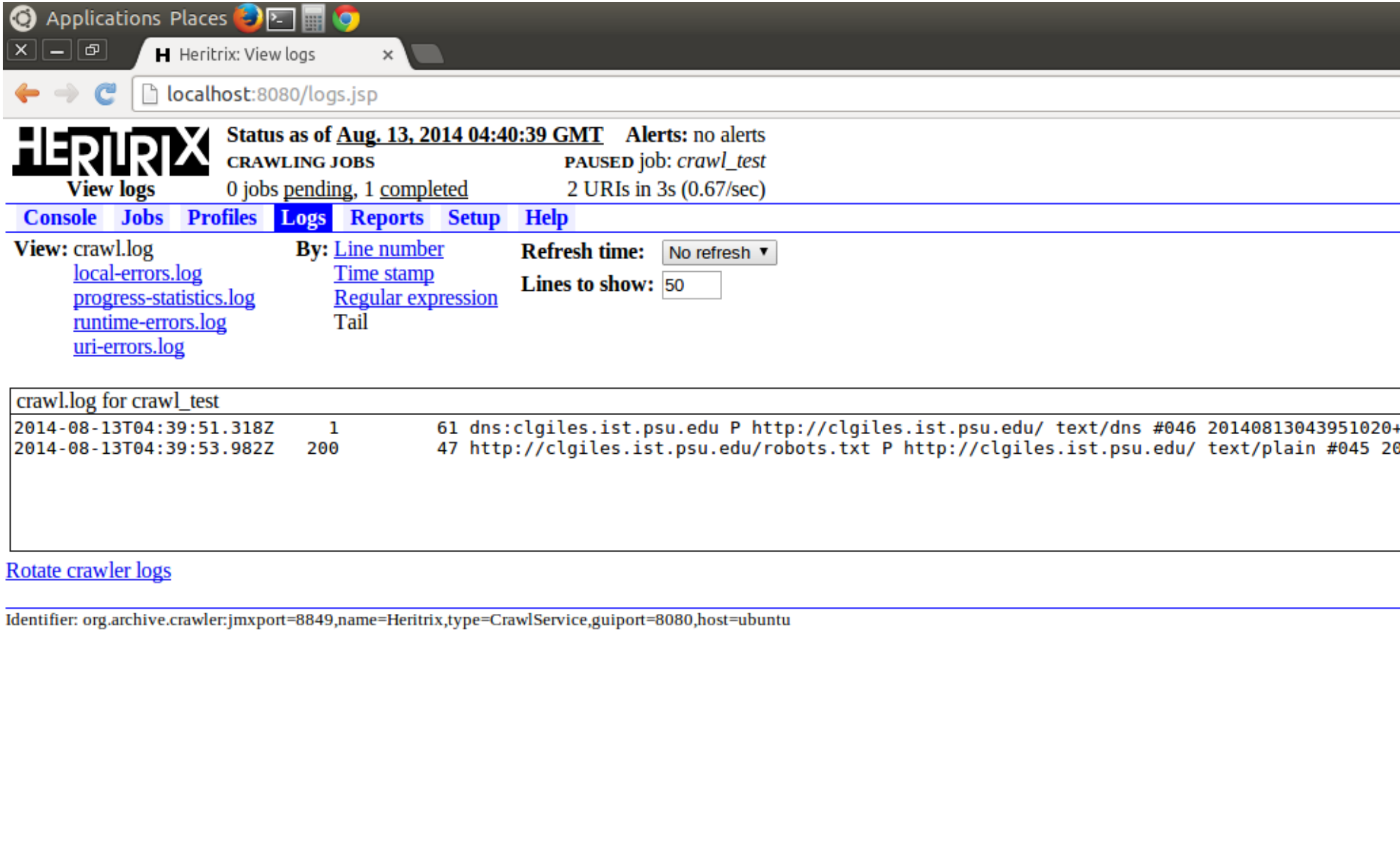

Applications Places
Heritrix: Crawl jobs
localhost:8080/jobs.jsp?message=Job%20created
HERITRIX
Crawl jobs
Status as of Aug. 13, 2014 04:39:14 GMT
Alerts: no alerts
HOLDING JOBS
1 jobs pending, 1 completed
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Job created
Create New Job
Based on existing job
Based on a recovery
Based on a profile
With defaults
Pending Jobs (1)
Name
Status
Options
crawl_test
Pending
View order
Edit configuration
Journal
Delete
Completed Jobs(1)
UID
Name
Status
Options
20140812043421791
default
Finished - Ended by operator
Crawl order
Crawl report
Seeds report
Seed file
Logs
Journal
Delete
Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=8080,host=ubuntu

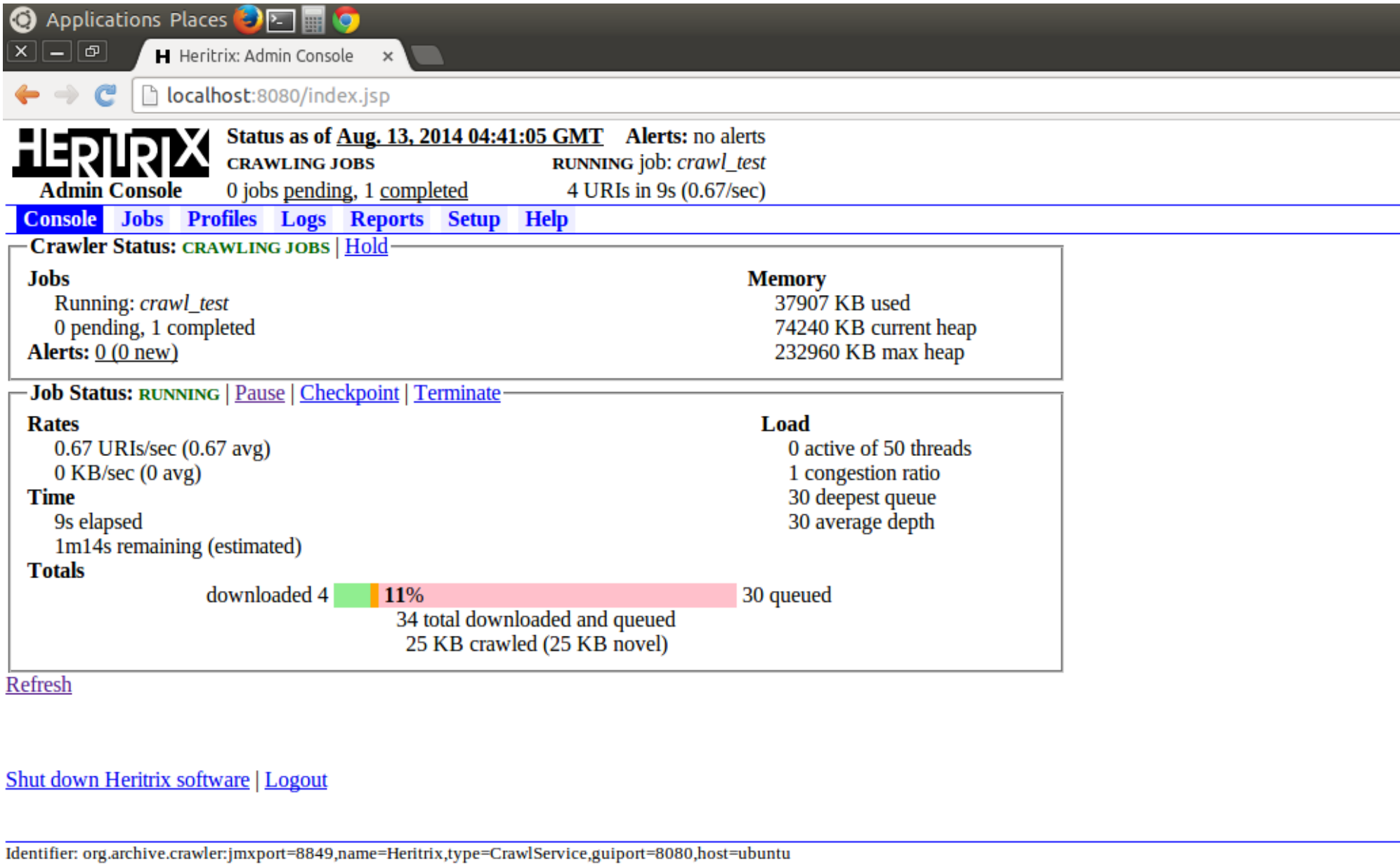

Applications Places
Heritrix: Admin Console
localhost:8080/index.jsp
HERITRIX
Admin Console
Status as of Aug. 13, 2014 04:39:54 GMT
Alerts: no alerts
CRAWLING JOBS
0 jobs pending, 1 completed
PAUSED job: crawl_test
2 URIs in 3s (0.67/sec)
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Crawler Status: CRAWLING JOBS | Hold
Jobs
Paused: crawl_test
0 pending, 1 completed
Alerts: 0 (0 new)
Memory
36329 KB used
74240 KB current heap
232960 KB max heap
Job Status: PAUSED | Resume | Checkpoint | Terminate
Rates
0.67 URIs/sec (0.67 avg)
0 KB/sec (0 avg)
Time
3s elapsed
1s remaining (estimated)
Load
0 active of 50 threads
1 congestion ratio
1 deepest queue
1 average depth
Totals
downloaded 2
66%
1 queued
3 total downloaded and queued
356 B crawled (356 B novel)
Paused Operations
View or Edit Frontier URIs
Refresh
Shut down Heritrix software | Logout

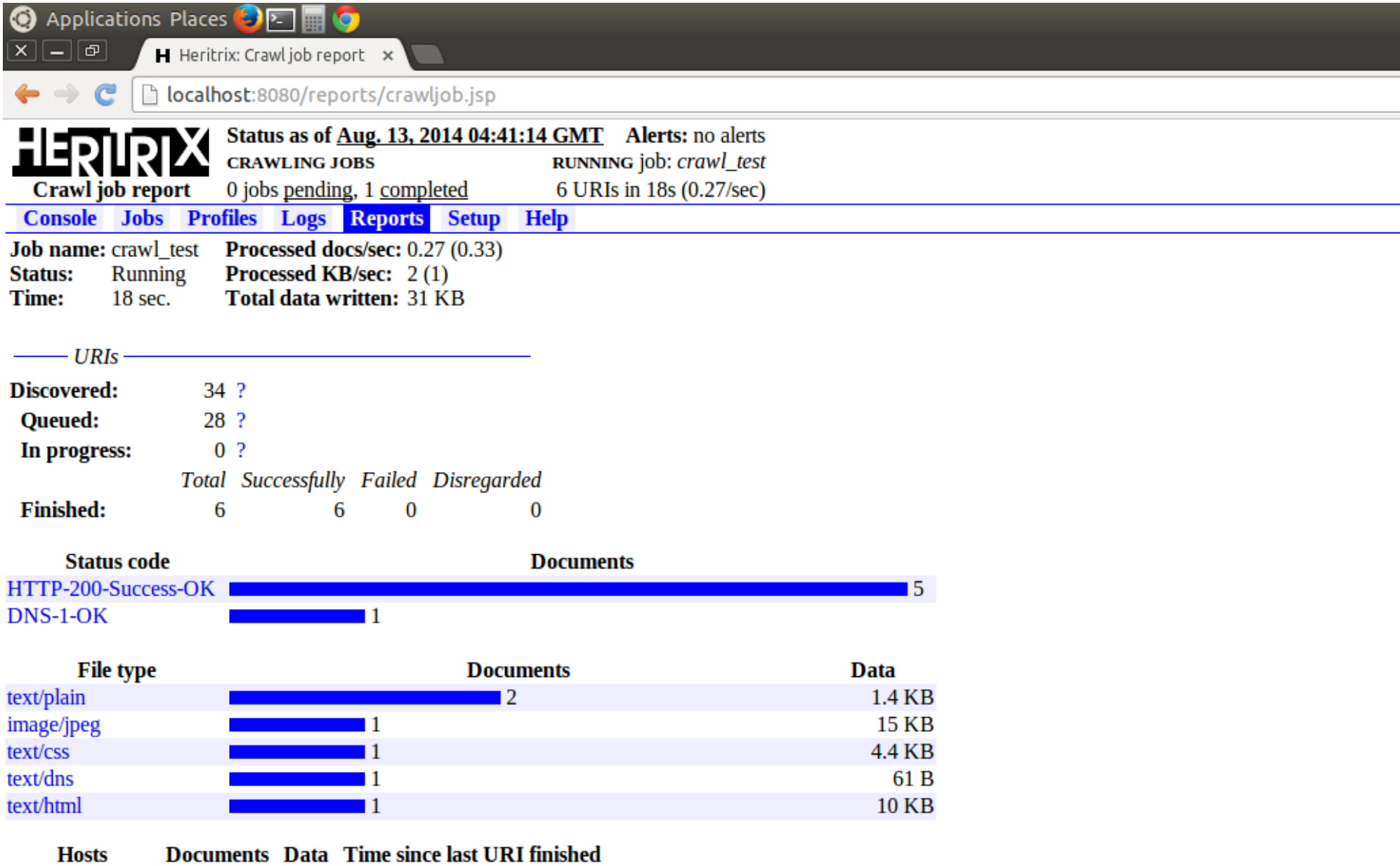

Applications Places
Heritrix: Admin Console
localhost:8080/index.jsp
HERITRIX
Admin Console
Status as of Aug. 13, 2014 04:39:54 GMT
Alerts: no alerts
CRAWLING JOBS
0 jobs pending, 1 completed
PAUSED job: crawl_test
2 URIs in 3s (0.67/sec)
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Crawler Status: CRAWLING JOBS | Hold
Jobs
Paused: crawl_test
0 pending, 1 completed
Alerts: 0 (0 new)
Memory
36329 KB used
74240 KB current heap
232960 KB max heap
Job Status: PAUSED | Resume | Checkpoint | Terminate
Rates
0.67 URIs/sec (0.67 avg)
0 KB/sec (0 avg)
Time
3s elapsed
1s remaining (estimated)
Load
0 active of 50 threads
1 congestion ratio
1 deepest queue
1 average depth
Totals
downloaded 2
66%
1 queued
3 total downloaded and queued
356 B crawled (356 B novel)
Paused Operations
View or Edit Frontier URIs
Refresh
Shut down Heritrix software | Logout

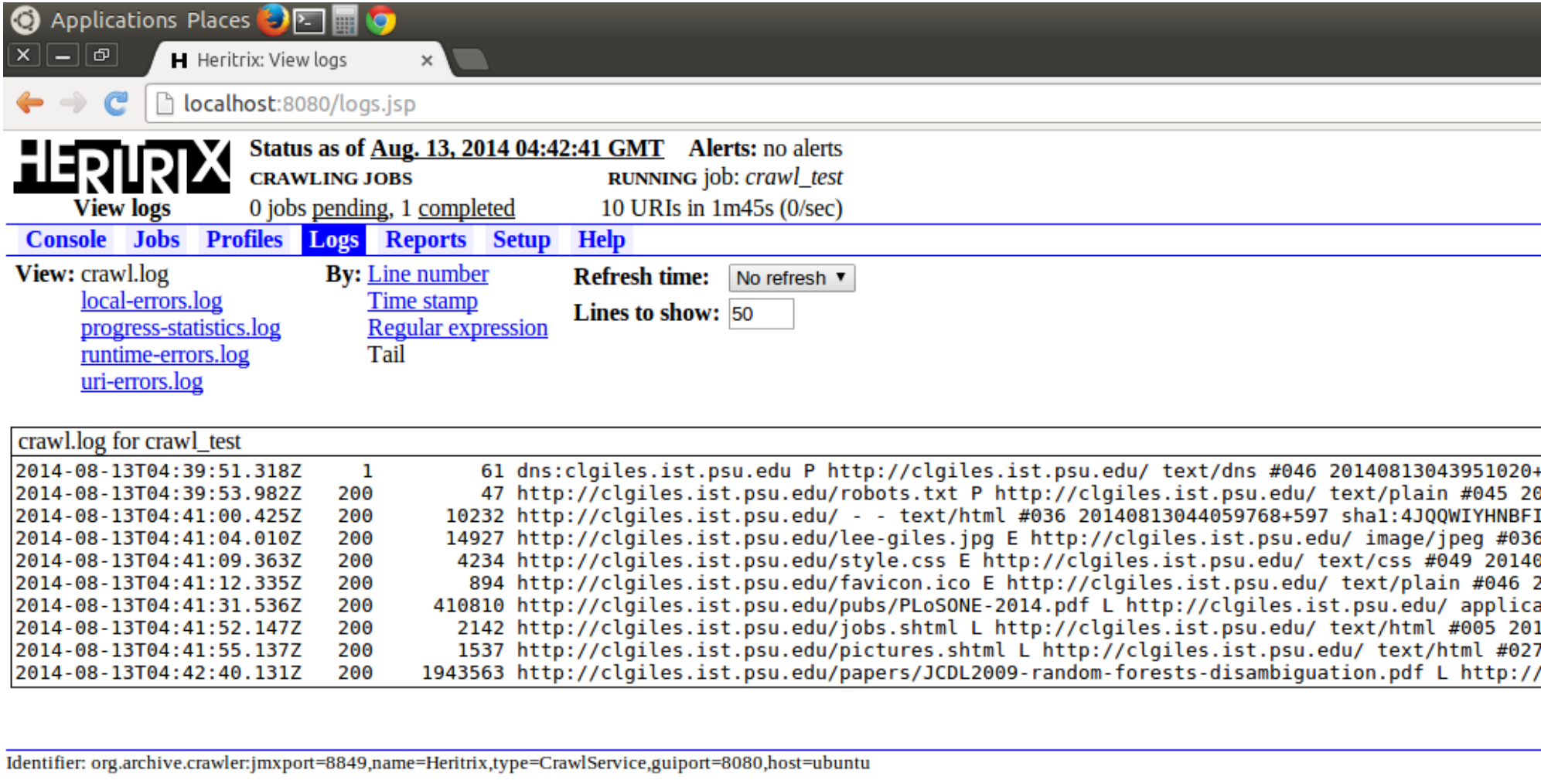

Applications Places
Heritrix: Reports
localhost:8080/reports.jsp
HERITRIX
Reports
Status as of Aug. 13, 2014 04:40:13 GMT
Alerts: no alerts
CRAWLING JOBS
PAUSED job: crawl_test
0 jobs pending, 1 completed
2 URIs in 3s (0.67/sec)
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Reports on ongoing crawl/current status
Crawl report
Seed report
Frontier report
1 queues: 1 active (0 in-process; 1 ready; 0 snoozed); 0 inactive; 0 retired; 0 exhausted
Processors report
ToeThread report
50 threads: 50 ABOUT_TO_GET_URI
The crawler generates reports when it finishes a job. Clicking here on Force generation of end-of-crawl Reports will force the writing of reports to disk. Clicking thi
for the generated reports. Each click overwrites previously generated reports. Use this facility when the crawler has hung threads that can't be interrupted.
Started crawl jobs (newest to oldest)
crawl_test Paused
default Crawl report Seed report Finished - Ended by operator
Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=8080,host=ubuntu

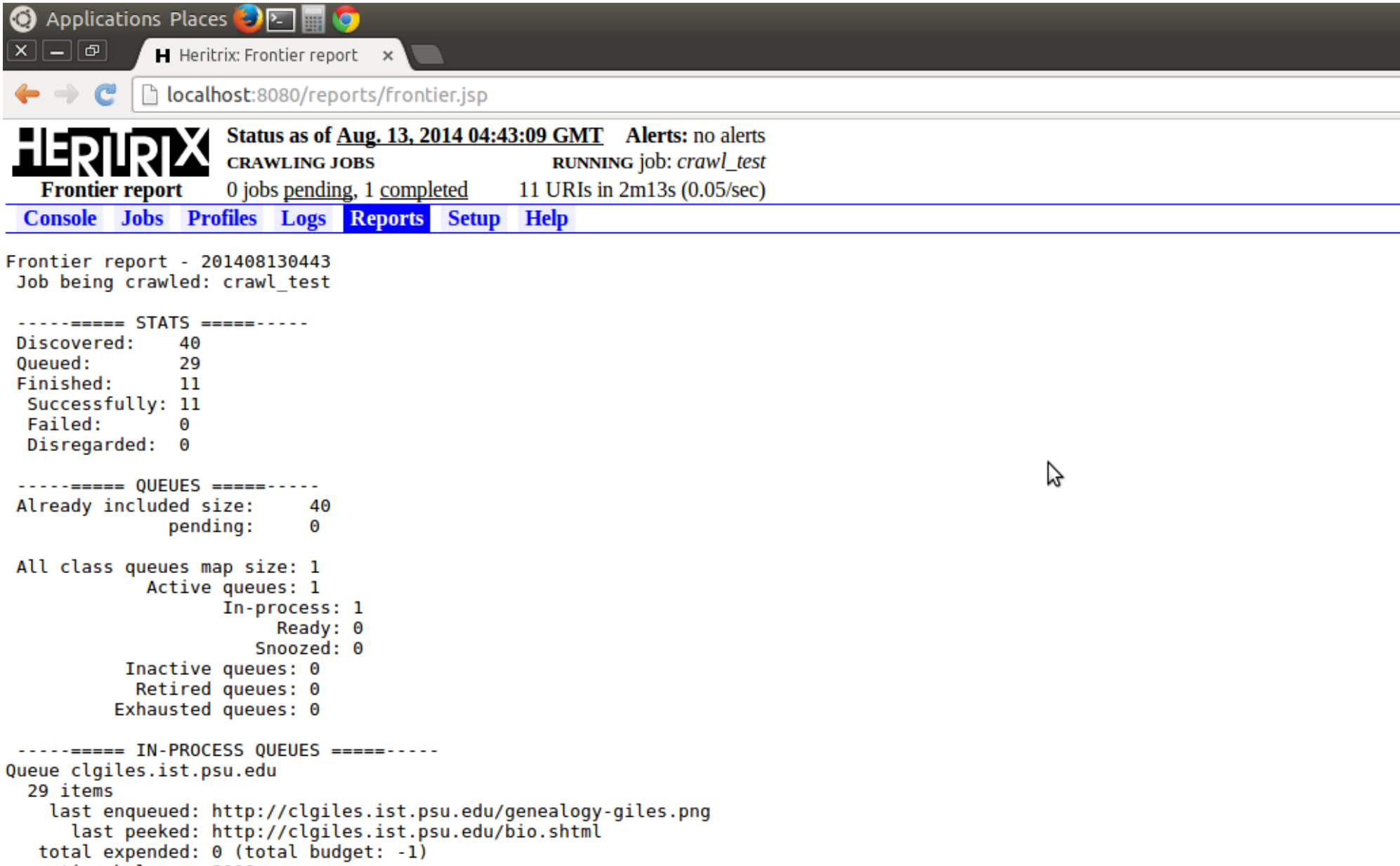

Applications Places
Heritrix: Frontier report
localhost:8080/reports/frontier.jsp
HERITRIX
Frontier report
Status as of Aug. 13, 2014 04:40:23 GMT
Alerts: no alerts
CRAWLING JOBS
PAUSED job: crawl_test
0 jobs pending, 1 completed
2 URIs in 3s (0.67/sec)
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Frontier report - 201408130440
Job being crawled: crawl_test
-----===== STATS =====-----
Discovered: 3
Queued: 1
Finished: 2
Successfully: 2
Failed: 0
Disregarded: 0
-----===== QUEUES =====-----
Already included size: 3
pending: 0
All class queues map size: 1
Active queues: 1
In-process: 0
Ready: 1
Snoozed: 0
Inactive queues: 0
Retired queues: 0
Exhausted queues: 0
-----===== IN-PROCESS QUEUES =====-----
-----===== READY QUEUES =====-----
Queue clgiles.ist.psu.edu
1 items
last enqueued: http://clgiles.ist.psu.edu/

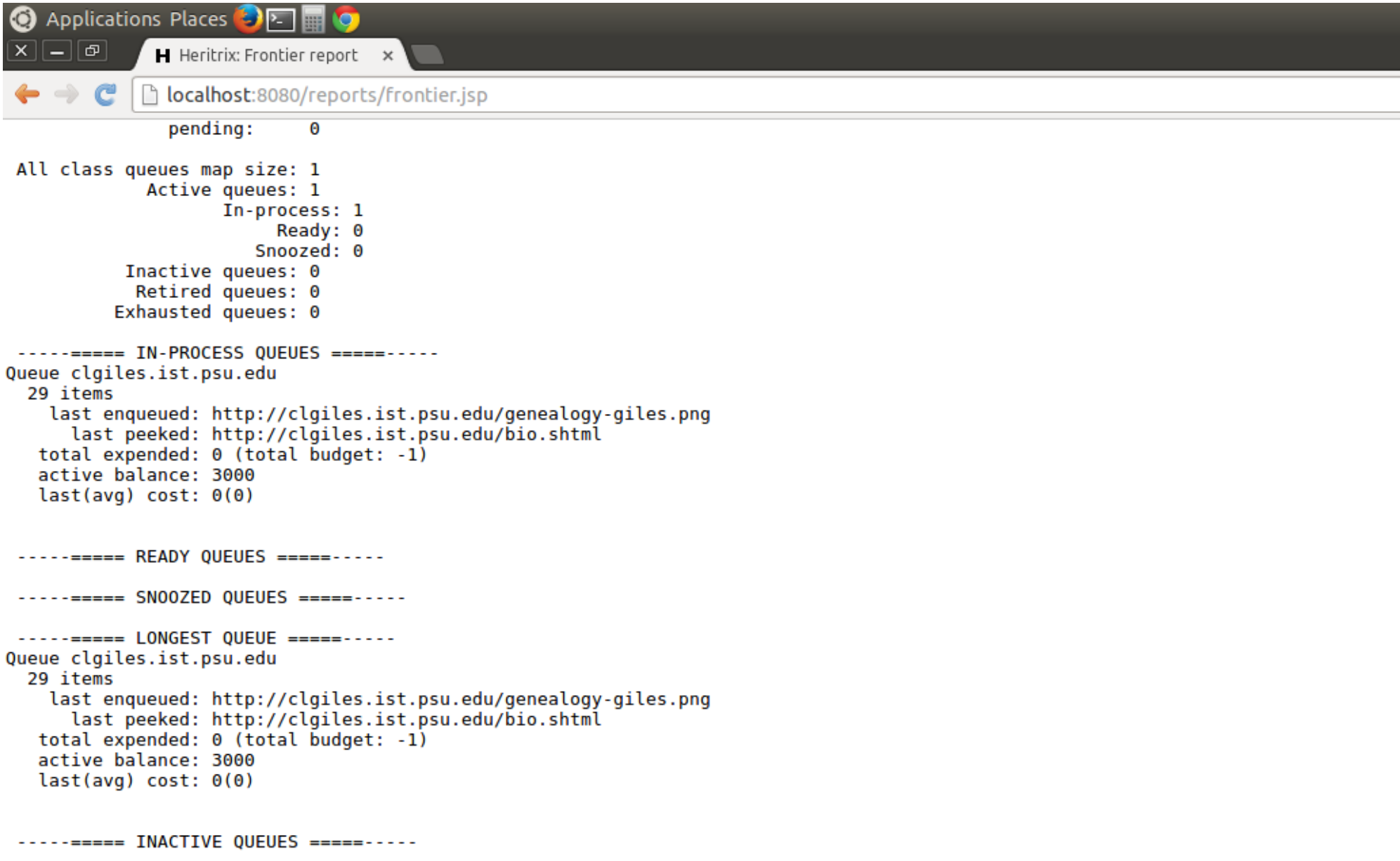

Applications Places
Heritrix: View logs
localhost:8080/logs.jsp
HERITRIX
View logs
Status as of Aug. 13, 2014 04:40:39 GMT
Alerts: no alerts
CRAWLING JOBS
PAUSED job: crawl_test
0 jobs pending, 1 completed
2 URIs in 3s (0.67/sec)
Console
Jobs
Profiles
Logs
Reports
Setup
Help
View: crawl.log
local-errors.log
progress-statistics.log
runtime-errors.log
uri-errors.log
By: Line number
Time stamp
Regular expression
Tail
Refresh time: No refresh
Lines to show: 50
crawl.log for crawl_test
2014-08-13T04:39:51.318Z 1 61 dns:clgiles.ist.psu.edu P http://clgiles.ist.psu.edu/ text/dns #046 20140813043951020+
2014-08-13T04:39:53.982Z 200 47 http://clgiles.ist.psu.edu/robots.txt P http://clgiles.ist.psu.edu/ text/plain #045 20
Rotate crawler logs
Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=8080,host=ubuntu


Applications Places
Heritrix: Admin Console
localhost:8080/index.jsp
HERITRIX
Admin Console
Status as of Aug. 13, 2014 04:41:05 GMT
Alerts: no alerts
CRAWLING JOBS
RUNNING job: crawl_test
0 jobs pending, 1 completed
4 URIs in 9s (0.67/sec)
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Crawler Status: CRAWLING JOBS | Hold
Jobs
Running: crawl_test
0 pending, 1 completed
Alerts: 0 (0 new)
Memory
37907 KB used
74240 KB current heap
232960 KB max heap
Job Status: RUNNING | Pause | Checkpoint | Terminate
Rates
0.67 URIs/sec (0.67 avg)
0 KB/sec (0 avg)
Time
9s elapsed
1m14s remaining (estimated)
Load
0 active of 50 threads
1 congestion ratio
30 deepest queue
30 average depth
Totals
downloaded 4
11%
30 queued
34 total downloaded and queued
25 KB crawled (25 KB novel)
Refresh
Shut down Heritrix software | Logout
Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=8080,host=ubuntu

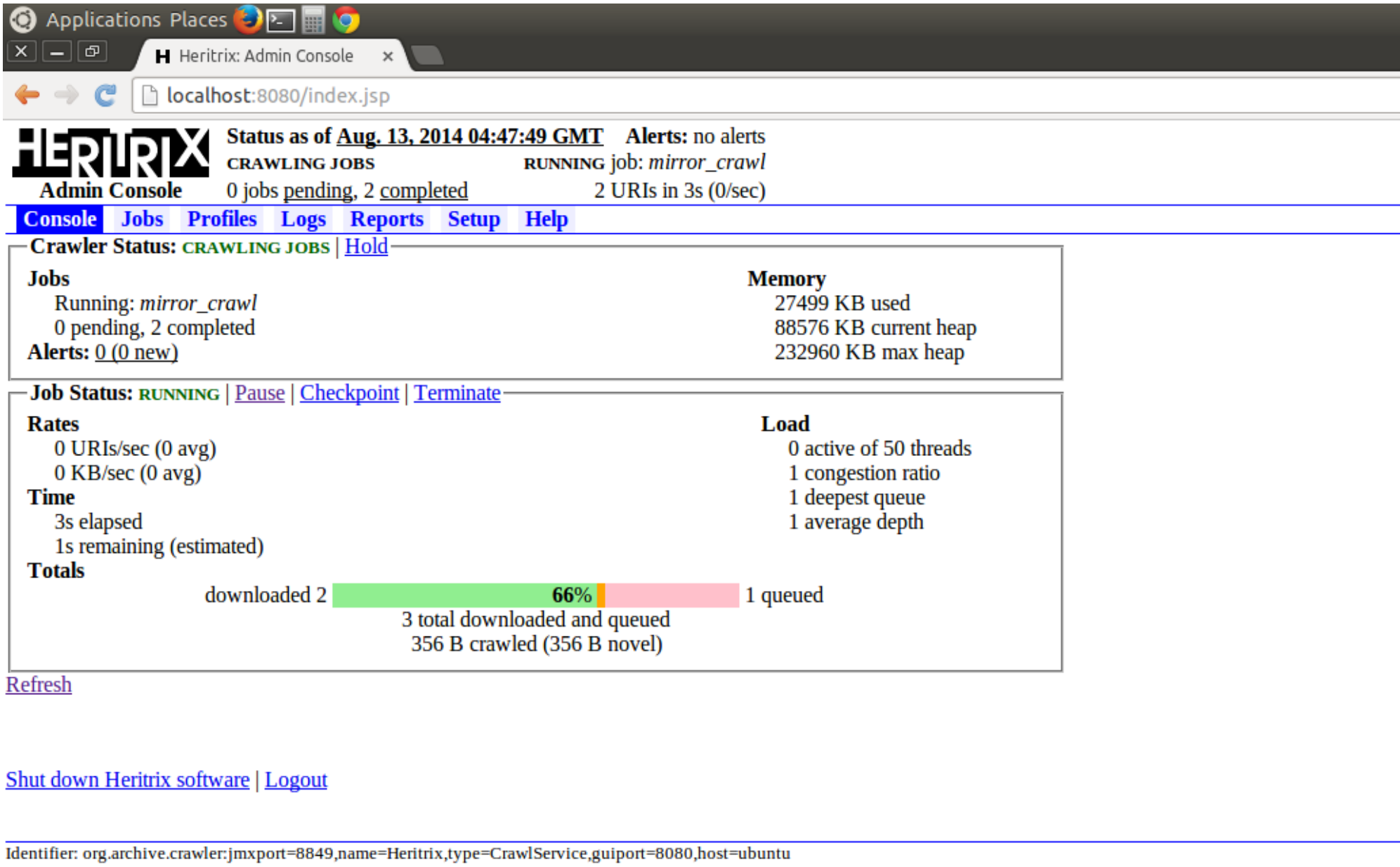

Applications Places
Heritrix: Crawl job report
localhost:8080/reports/crawljob.jsp
HERITRIX
Crawl job report
Status as of Aug. 13, 2014 04:41:14 GMT
Alerts: no alerts
CRAWLING JOBS
0 jobs pending, 1 completed
RUNNING job: crawl_test
6 URIs in 18s (0.27/sec)
Console Jobs Profiles Logs Reports Setup Help
Job name: crawl_test
Status: Running
Time: 18 sec.
Processed docs/sec: 0.27 (0.33)
Processed KB/sec: 2 (1)
Total data written: 31 KB
URIs
Discovered: 34 ?
Queued: 28 ?
In progress: 0 ?
Total Successfully Failed Disregarded
Finished: 6 6 0 0
Status code
Documents
HTTP-200-Success-OK 5
DNS-1-OK 1
File type
Documents
Data
text/plain 2 1.4 KB
image/jpeg 1 15 KB
text/css 1 4.4 KB
text/dns 1 61 B
text/html 1 10 KB
Hosts
Documents
Data
Time since last URI finished

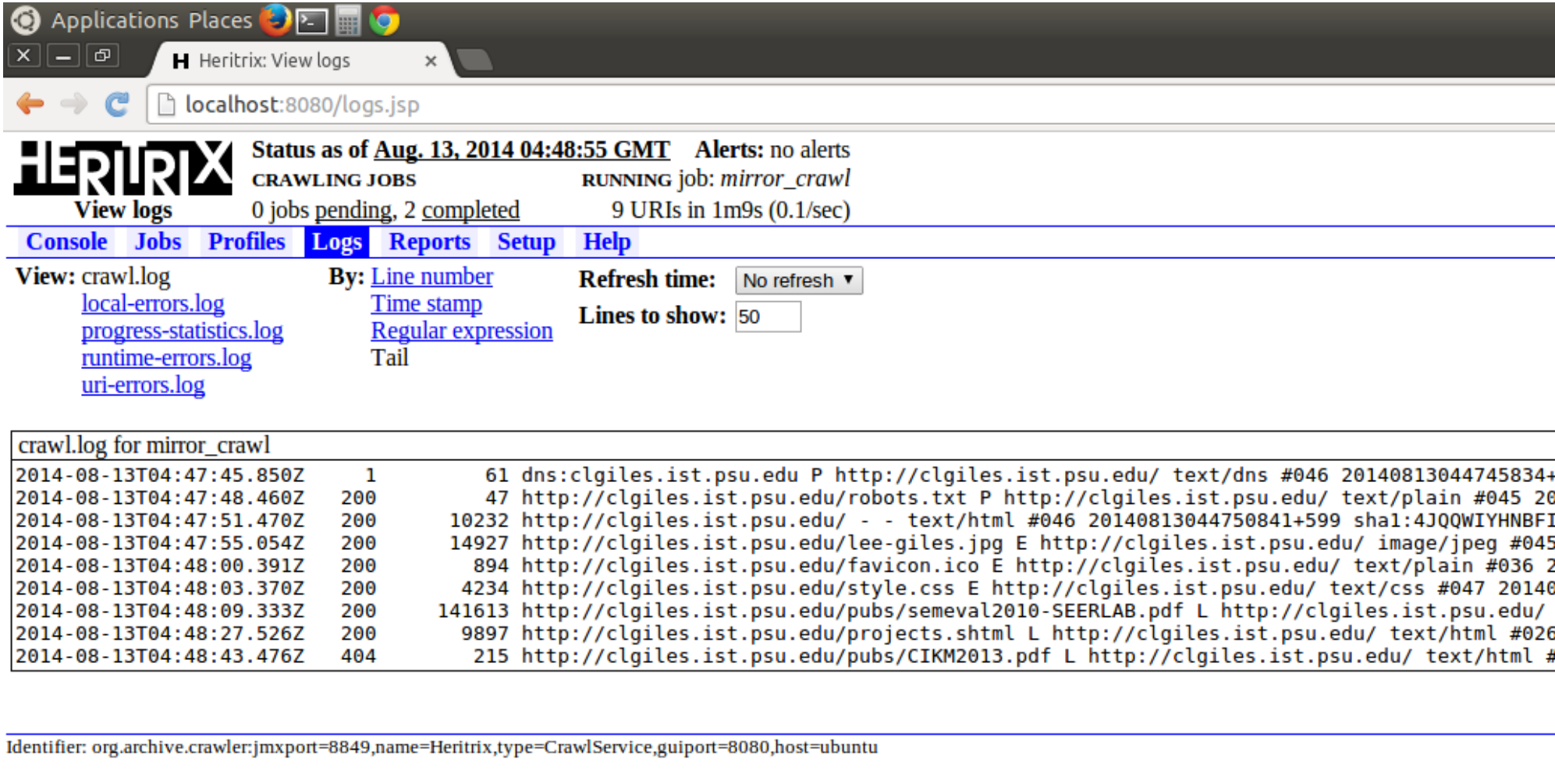

Applications Places
Heritrix: View logs
localhost:8080/logs.jsp
HERITRIX
View logs
Status as of Aug. 13, 2014 04:42:41 GMT Alerts: no alerts
CRAWLING JOBS
RUNNING job: crawl_test
0 jobs pending, 1 completed
10 URIs in 1m45s (0/sec)
Console Jobs Profiles Logs Reports Setup Help
View: crawl.log
local-errors.log
progress-statistics.log
runtime-errors.log
uri-errors.log
By: Line number
Time stamp
Regular expression
Tail
Refresh time: No refresh
Lines to show: 50
crawl.log for crawl_test
2014-08-13T04:39:51.318Z 1 61 dns:clgiles.ist.psu.edu P http://clgiles.ist.psu.edu/ text/dns #046 20140813043951020+
2014-08-13T04:39:53.982Z 200 47 http://clgiles.ist.psu.edu/robots.txt P http://clgiles.ist.psu.edu/ text/plain #045 20
2014-08-13T04:41:00.425Z 200 10232 http://clgiles.ist.psu.edu/ - - text/html #036 20140813044059768+597 sha1:4JQQWIYHNBFI
2014-08-13T04:41:04.010Z 200 14927 http://clgiles.ist.psu.edu/lee-giles.jpg E http://clgiles.ist.psu.edu/ image/jpeg #036
2014-08-13T04:41:09.363Z 200 4234 http://clgiles.ist.psu.edu/style.css E http://clgiles.ist.psu.edu/ text/css #049 20140
2014-08-13T04:41:12.335Z 200 894 http://clgiles.ist.psu.edu/favicon.ico E http://clgiles.ist.psu.edu/ text/plain #046 2
2014-08-13T04:41:31.536Z 200 410810 http://clgiles.ist.psu.edu/pubs/PLoSONE-2014.pdf L http://clgiles.ist.psu.edu/ applica
2014-08-13T04:41:52.147Z 200 2142 http://clgiles.ist.psu.edu/jobs.shtml L http://clgiles.ist.psu.edu/ text/html #005 201
2014-08-13T04:41:55.137Z 200 1537 http://clgiles.ist.psu.edu/pictures.shtml L http://clgiles.ist.psu.edu/ text/html #027
2014-08-13T04:42:40.131Z 200 1943563 http://clgiles.ist.psu.edu/papers/JCDL2009-random-forests-disambiguation.pdf L http://
Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=8080,host=ubuntu


Applications Places
Heritrix: Frontier report
localhost:8080/reports/frontier.jsp
HERITRIX
Frontier report
Status as of Aug. 13, 2014 04:43:09 GMT Alerts: no alerts
CRAWLING JOBS RUNNING job: crawl_test
0 jobs pending, 1 completed 11 URIs in 2m13s (0.05/sec)
Console Jobs Profiles Logs Reports Setup Help
Frontier report - 201408130443
Job being crawled: crawl_test
-----===== STATS =====-----
Discovered: 40
Queued: 29
Finished: 11
Successfully: 11
Failed: 0
Disregarded: 0
-----===== QUEUES =====-----
Already included size: 40
pending: 0
All class queues map size: 1
Active queues: 1
In-process: 1
Ready: 0
Snoozed: 0
Inactive queues: 0
Retired queues: 0
Exhausted queues: 0
-----===== IN-PROCESS QUEUES =====-----
Queue clgiles.ist.psu.edu
29 items
last enqueued: http://clgiles.ist.psu.edu/genealogy-giles.png
last peeked: http://clgiles.ist.psu.edu/bio.shtml
total expended: 0 (total budget: -1)


Applications Places
Heritrix: Frontier report
localhost:8080/reports/frontier.jsp
pending: 0
All class queues map size: 1
Active queues: 1
In-process: 1
Ready: 0
Snoozed: 0
Inactive queues: 0
Retired queues: 0
Exhausted queues: 0
-----===== IN-PROCESS QUEUES =====-----
Queue clgiles.ist.psu.edu
29 items
last enqueued: http://clgiles.ist.psu.edu/genealogy-giles.png
last peeked: http://clgiles.ist.psu.edu/bio.shtml
total expended: 0 (total budget: -1)
active balance: 3000
last(avg) cost: 0(0)
-----===== READY QUEUES =====-----
-----===== SNOOZED QUEUES =====-----
-----===== LONGEST QUEUE =====-----
Queue clgiles.ist.psu.edu
29 items
last enqueued: http://clgiles.ist.psu.edu/genealogy-giles.png
last peeked: http://clgiles.ist.psu.edu/bio.shtml
total expended: 0 (total budget: -1)
active balance: 3000
last(avg) cost: 0(0)
-----===== INACTIVE QUEUES =====-----


Applications Places
1:13 PM
sdas
sdas@ubuntu: ~/setups/heritrix-1.14.3/jobs/crawl_test-20140813043422142/arcs
File Edit View Search Terminal Help
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/crawl_test-20140813043422142$ pwd
/home/sdas/setups/heritrix-1.14.3/jobs/crawl_test-20140813043422142
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/crawl_test-20140813043422142$ ls
arcs frontier-report.txt order.xml seeds-report.txt
checkpoints hosts-report.txt processors-report.txt seeds.txt
crawl-manifest.txt logs responsecode-report.txt state
crawl-report.txt mimetype-report.txt scratch state.job
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/crawl_test-20140813043422142$ cd arcs/
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/crawl_test-20140813043422142/arcs$ ls
IAH-20140813043951-00000-ubuntu.arc.gz
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/crawl_test-20140813043422142/arcs$
sdas@ubuntu: ~/setup...


Applications Places
Heritrix: Admin Console
localhost:8080/index.jsp
HERITRIX
Admin Console
Status as of Aug. 13, 2014 04:47:49 GMT
Alerts: no alerts
CRAWLING JOBS
RUNNING job: mirror_crawl
0 jobs pending, 2 completed
2 URIs in 3s (0/sec)
Console
Jobs
Profiles
Logs
Reports
Setup
Help
Crawler Status: CRAWLING JOBS | Hold
Jobs
Running: mirror_crawl
0 pending, 2 completed
Alerts: 0 (0 new)
Memory
27499 KB used
88576 KB current heap
232960 KB max heap
Job Status: RUNNING | Pause | Checkpoint | Terminate
Rates
0 URIs/sec (0 avg)
0 KB/sec (0 avg)
Time
3s elapsed
1s remaining (estimated)
Load
0 active of 50 threads
1 congestion ratio
1 deepest queue
1 average depth
Totals
downloaded 2
66%
1 queued
3 total downloaded and queued
356 B crawled (356 B novel)
Refresh
Shut down Heritrix software | Logout
Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=8080,host=ubuntu


Applications Places
Heritrix: View logs
localhost:8080/logs.jsp
HERITRIX
View logs
Status as of Aug. 13, 2014 04:48:55 GMT Alerts: no alerts
CRAWLING JOBS
0 jobs pending, 2 completed
RUNNING job: mirror_crawl
9 URIs in 1m9s (0.1/sec)
Console Jobs Profiles Logs Reports Setup Help
View: crawl.log
local-errors.log
progress-statistics.log
runtime-errors.log
uri-errors.log
By: Line number
Time stamp
Regular expression
Tail
Refresh time: No refresh
Lines to show: 50
crawl.log for mirror_crawl
2014-08-13T04:47:45.850Z 1 61 dns:clgiles.ist.psu.edu P http://clgiles.ist.psu.edu/ text/dns #046 20140813044745834+
2014-08-13T04:47:48.460Z 200 47 http://clgiles.ist.psu.edu/robots.txt P http://clgiles.ist.psu.edu/ text/plain #045 20
2014-08-13T04:47:51.470Z 200 10232 http://clgiles.ist.psu.edu/ - - text/html #046 20140813044750841+599 sha1:4JQQWIYHNBFI
2014-08-13T04:47:55.054Z 200 14927 http://clgiles.ist.psu.edu/lee-giles.jpg E http://clgiles.ist.psu.edu/ image/jpeg #045
2014-08-13T04:48:00.391Z 200 894 http://clgiles.ist.psu.edu/favicon.ico E http://clgiles.ist.psu.edu/ text/plain #036 2
2014-08-13T04:48:03.370Z 200 4234 http://clgiles.ist.psu.edu/style.css E http://clgiles.ist.psu.edu/ text/css #047 20140
2014-08-13T04:48:09.333Z 200 141613 http://clgiles.ist.psu.edu/pubs/semeval2010-SEERLAB.pdf L http://clgiles.ist.psu.edu/
2014-08-13T04:48:27.526Z 200 9897 http://clgiles.ist.psu.edu/projects.shtml L http://clgiles.ist.psu.edu/ text/html #026
2014-08-13T04:48:43.476Z 404 215 http://clgiles.ist.psu.edu/pubs/CIKM2013.pdf L http://clgiles.ist.psu.edu/ text/html #
Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=8080,host=ubuntu


Applications Places
1:14 PM sdas
sdas@ubuntu: ~/setups/heritrix-1.14.3/jobs/mirror_crawl-20140813044606002/mirror/clgiles.ist.psu.edu/pubs
File Edit View Search Terminal Help
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/mirror_crawl-20140813044606002$ pwd
/home/sdas/setups/heritrix-1.14.3/jobs/mirror_crawl-20140813044606002
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/mirror_crawl-20140813044606002$ ls
checkpoints hosts-report.txt order.xml seeds-report.txt
crawl-manifest.txt logs processors-report.txt seeds.txt
crawl-report.txt mimetype-report.txt responsecode-report.txt state
frontier-report.txt mirror scratch state.job
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/mirror_crawl-20140813044606002$ cd mirror/
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/mirror_crawl-20140813044606002/mirror$ ls
clgiles.ist.psu.edu
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/mirror_crawl-20140813044606002/mirror$ cd clgile
s.ist.psu.edu/
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/mirror_crawl-20140813044606002/mirror/clgiles.is
t.psu.edu$ ls
collaborators.shtml index.html projects.shtml robots.txt
favicon.ico lee-giles.jpg pubs style.css
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/mirror_crawl-20140813044606002/mirror/clgiles.is
t.psu.edu$ cd pubs/
sdas@ubuntu:~/setups/heritrix-1.14.3/jobs/mirror_crawl-20140813044606002/mirror/clgiles.is
t.psu.edu/pubs$ ls
AAAI2012-name-ethnicity.pdf semeval2010-SEERLAB.pdf
sdas@ubuntu: ~/setup...