ArchiveBot Documentation

Release 1.6

ArchiveTeam

Jul 22, 2017

Contents

1 Commands **3**

1.1 archive . . . 3
1.2 abort . . . 5
1.3 archiveonly . . . 5
1.4 explain . . . 6
1.5 archiveonly < FILE . . . 7
1.6 ignore . . . 7
1.7 unignore . . . 8
1.8 ignoreset . . . 8
1.9 ignorereports . . . 9
1.10 delay . . . 9
1.11 concurrency . . . 9
1.12 yahoo . . . 9
1.13 expire . . . 10
1.14 status . . . 10
1.15 pending . . . 11
1.16 whereis . . . 11

2 Indices and tables **13**

Homepage http://www.archiveteam.org/index.php?title=ArchiveBot

Contents:

CHAPTER 1

Commands

ArchiveBot listens to commands prefixed with `!`.

archive

`!archive URL`, `!a URL` begin recursive retrieval from a URL

```
> !archive http://artscene.textfiles.com/litpacks/
< Archiving http://artscene.textfiles.com/litpacks/.
< Use !status 43z7a11vo6of3a7i173441dtc for updates, !abort
  43z7a11vo6of3a7i173441dtc to abort.
```

ArchiveBot does not ascend to parent links. This means that everything under the `litpacks` directory will be downloaded. For example, `/litpacks/hello.html` will be downloaded but not `/hello.html`.

If you leave out the trailing slash, eg `/litpacks`, it will consider that to be a file and download everything under `/`.

URLs are treated as case-sensitive. `/litpacks` is different from `/LitPacks`.

Accepted parameters

`--ignore-sets SET1,...,SETN` specify sets of URL patterns to ignore:

```
> !archive http://example.blogspot.com/ncr --ignore-sets=blogs,forums
< Archiving http://example.blogspot.com/ncr.
< 14 ignore patterns loaded.
< Use !status 5sid4pgxkiu6zynhbt3q1gi2s for updates, !abort
  5sid4pgxkiu6zynhbt3q1gi2s to abort.
```

Known sets are listed in db/ignore_patterns/.

Aliases: `--ignoresets`, `--ignore_sets`, `--ignoreset`, `--ignore-set`, `--ignore_set`, `--ig-set`, `--igset`

--no-offsite-links do not download links to offsite pages:

```
> !archive http://example.blogspot.com/ncr
>    --no-offsite-links
< Archiving http://example.blogspot.com/ncr.
< Offsite links will not be grabbed.
< Use !status 5sid4pgxkiu6zynhbt3q1gi2s for updates, !abort
  5sid4pgxkiu6zynhbt3q1gi2s to abort.
```

ArchiveBot's default behavior with !archive is to recursively fetch all pages that are descendants of the starting URL, as well as all linked pages and their requisites. This is often useful for preserving a page's context in time. However, this can sometimes result in an undesirably large archive. Specifying `--no-offsite-links` preserves recursive retrieval but does not follow links to offsite hosts.

Please note that ArchiveBot considers `www.example.com` and `example.com` to be different hosts, so if you have a website that uses both, you should not specify `--no-offsite-links`.

Aliases: `--nooffsitelinks`, `--no-offsite`, `--nooffsite`

--user-agent-alias ALIAS specify a user-agent to use:

```
> !archive http://artscene.textfiles.com/litpacks/
    --user-agent-alias=firefox
< Archiving http://artscene.textfiles.com/litpacks/.
< Using user-agent Mozilla/5.0 (Windows NT 5.1; rv:31.0)
  Gecko/20100101 Firefox/31.0.
< Use !status 43z7a11vo6of3a7i173441dtc for updates, !abort
  43z7a11vo6of3a7i173441dtc to abort.
```

This option makes the job present the given user-agent. It can be useful for archiving sites that (still) do user-agent detection.

See db/user_agents for a list of recognized aliases.

Aliases: `--useragentalias`, `--user-agent`, `--useragent`

--pipeline TAG specify pipeline to use:

```
> !archive http://example.blogspot.com/ncr
    --pipeline=superfast
< Archiving http://example.blogspot.com/ncr.
< Job will run on a pipeline whose name contains "superfast".
< Use !status 5sid4pgxkiu6zynhbt3q1gi2s for updates, !abort
  5sid4pgxkiu6zynhbt3q1gi2s to abort.
```

Pipeline operators assign nicknames to pipelines. Oftentimes, these nicknames describe the pipeline: datacenter, special modifications, etc. This option can be used to load jobs onto those pipelines.

In the above example, the following pipeline nicks would match the given tag:

- superfast
- ovhca1-superfast-47

--phantomjs access pages via PhantomJS

--phantomjs-wait set number of seconds between PhantomJS requests; defaults to 2.0

--phantomjs-scroll maximum number of times to scroll a page in PhantomJS; defaults to 100

--no-phantomjs-smart-scroll disable PhantomJS' end-of-page detection and always scroll `--phantomjs-scroll` number of times; off by default

PhantomJS mode is enabled if any of the `--*phantomjs*` options are passed.

`--explain` alias for `!explain` adds a short note explaining the purpose of the archiving job

`--delay` alias for `!delay` (in milliseconds) only allows a single value; to provide a range, use `!delay`

`--concurrency` alias for `!concurrency` sets number of workers for job (use with care!)

abort

`!abort IDENT` abort a job:

```
> !abort 1q2qydhkeh3gfnrcxuf6py70b
< Initiating abort for job 1q2qydhkeh3gfnrcxuf6py70b.
```

archiveonly

`!archiveonly URL`, **`!ao URL`** non-recursive retrieval of the given URL:

```
> !archiveonly http://store.steampowered.com/livingroom
< Archiving http://store.steampowered.com/livingroom without
  recursion.
> Use !status 1q2qydhkeh3gfnrcxuf6py70b for updates, !abort
  1q2qydhkeh3gfnrcxuf6py70b to abort.
```

Accepted parameters

`--ignore-sets SET1,...,SETN` specify sets of URL patterns to ignore:

```
> !archiveonly http://example.blogspot.com/ --ignore-sets=blogs,forums
< Archiving http://example.blogspot.com/ without recursion.
< 14 ignore patterns loaded.
< Use !status 5sid4pgxkiu6zynhbt3q1gi2s for updates, !abort
  5sid4pgxkiu6zynhbt3q1gi2s to abort.
```

Known sets are listed in db/ignore_patterns/.

`--user-agent-alias ALIAS` specify a user-agent to use:

```
> !archiveonly http://artscene.textfiles.com/litpacks/
    --user-agent-alias=firefox
< Archiving http://artscene.textfiles.com/litpacks/ without
  recursion.
< Using user-agent Mozilla/5.0 (Windows NT 5.1; rv:31.0)
  Gecko/20100101 Firefox/31.0.
< Use !status 43z7a11vo6of3a7i173441dtc for updates, !abort
  43z7a11vo6of3a7i173441dtc to abort.
```

This option makes the job present the given user-agent. It can be useful for archiving sites that (still) do user-agent detection. See db/user_agents for a list of recognized aliases.

`--pipeline TAG` specify pipeline to use:

```
> !archiveonly http://example.blogspot.com/
    --pipeline=superfast
< Archiving http://example.blogspot.com/.
< Job will run on a pipeline whose name contains "superfast".
< Use !status 5sid4pgxkiu6zynhbt3q1gi2s for updates, !abort
  5sid4pgxkiu6zynhbt3q1gi2s to abort.
```

--youtube-dl

> **Warning:** This is a new feature; not all pipelines support it. To find a pipeline that supports youtube-dl, use the ArchiveBot pipeline monitor page and look for a pipeline whose version is newer than 20150512.01.

attempt to download videos using youtube-dl (experimental):

```
> !archiveonly https://example.website/fun-video-38214 --youtube-dl
< Queued https://example.website/fun-video-38214 for archival without
  recursion.
< Options: youtube-dl: yes
< Use !status dma5g7xcy0r3gbmisqshkpkoe for updates, !abort
  dma5g7xcy0r3gbmisqshkpkoe to abort.
```

When –youtube-dl is passed, ArchiveBot will attempt to download videos embedded in HTML pages it encounters in the crawl using youtube-dl (http://rg3.github.io/youtube-dl/). youtube-dl can recognize many different embedding formats, but success is not guaranteed.

If you are going to use this option, please watch your job's progress on the dashboard. If you see MP4 or WebM files in the download log, your videos were probably saved. (You can click on links in the download log to confirm.)

Video playback is not yet well-supported in web archive playback tools. As of May 2015:

- pywb v0.9 (https://github.com/ikreymer/pywb) is known to work.
- https://github.com/ikreymer/webarchiveplayer is based on pywb 0.8, and might work.
- The Internet Archive's Wayback Machine does not present videos in ArchiveBot WARCs. (Wayback may not support the record convention used by ArchiveBot and/or may not support video playback at all.)

--phantomjs access pages via PhantomJS

--phantomjs-wait set number of seconds between PhantomJS requests; defaults to 2.0

--phantomjs-scroll maximum number of times to scroll a page in PhantomJS; defaults to 100

--no-phantomjs-smart-scroll disable PhantomJS' end-of-page detection and always scroll `--phantomjs-scroll` number of times; off by default

PhantomJS mode is enabled if any of the `--*phantomjs*` options are passed.

explain

!explain IDENT NOTE, !ex IDENT NOTE add a short note to explain why this site is being archived:

```
> !explain byu50bzfdbnlyl6mrgn6dd24h shutting down 7/31
> Added note "shutting down 7/31" to job byu50bzfdbnlyl6mrgn6dd24h.
```

Pipeline operators (really, anyone) may want to know why a job is running. This becomes particularly important when a job grows very large (hundreds of gigabytes). While this can be done via IRC, IRC communication is asynchronous, people can be impatient, and a rationale can usually be summed up very concisely.

archiveonly < FILE

!archiveonly < URL, **!ao < URL** archive each URL in the text file at URL:

```
> !archiveonly < https://www.example.com/some-file.txt
< Archiving URLs in https://www.example.com/some-file.txt without
  recursion.
> Use !status byu50bzfdbnlyl6mrgn6dd24h for updates, !abort
  byu50bzfdbnlyl6mrgn6dd24h to abort.
```

The text file should list one URL per line. Both UNIX and Windows line endings are accepted.

Accepted parameters

`!archiveonly < URL` accepts the same parameters as `!archiveonly`. A quick reference:

--ignore-sets SET1,...,SETN specify sets of URL patterns to ignore

--user-agent-alias ALIAS specify a user-agent to use

--pipeline TAG specify pipeline to use

--youtube-dl attempt to download videos using youtube-dl

--phantomjs access pages via PhantomJS

--phantomjs-wait set number of seconds between PhantomJS requests; defaults to 2.0

--phantomjs-scroll maximum number of times to scroll a page in PhantomJS; defaults to 100

--no-phantomjs-smart-scroll disable PhantomJS' end-of-page detection and always scroll `--phantomjs-scroll` number of times; off by default

ignore

!ignore IDENT PATTERN, **!ig IDENT PATTERN** add an ignore pattern:

```
> !ig 1q2qydhkeh3gfnrcxuf6py70b obnoxious\?foo=\d+
< Added ignore pattern obnoxious\?foo=\d+ to job
  1q2qydhkeh3gfnrcxuf6py70b.
```

The pattern must be expressed as regular expressions. For more information, see:

- http://docs.python.org/3/howto/regex.html#regex-howto
- http://docs.python.org/3/library/re.html#regular-expression-syntax

Two strings, `{primary_url}` and `{primary_netloc}`, have special meaning.

`{primary_url}` expands to the top-level URL. For `!archive` jobs, this is the initial URL. For `!archiveonly < FILE` jobs, `{primary_url}` is the top-level URL that owns the descendant being archived.

`{primary_netloc}` is the auth/host/port section of `{primary_url}`.

Examples

1. To ignore everything on domain1.com and its subdomains, use pattern `^https?://([^/]+\.)?domain1\.com/`
2. To ignore everything *except* URLs on domain1.com or domain2.com, use pattern `^(?!https?://(domain1\.com|domain2\.com)/)`
3. To keep subdomains on domain1.com as well, use pattern `^(?!https?://(([^/]+\.)?domain1\.com|domain2\.com)/)`
4. For `!archive` jobs on subdomain blogs (such as Tumblr), the following pattern ignores all URLs except the initial URL, sub-URLs of the initial URL, and media/asset servers: `^http://(?!({primary_netloc}|\d+\.media\.example\.com|assets\.example\.com)).*`
5. Say you have this URL file:

   ```
   http://www.example.com/foo.html
   http://www.bar.org:8080/qux.html
   ```

   and you submit it as an `!archiveonly < FILE` job.

   When retrieving requisites of `http://www.example.com/foo.html`, `{primary_url}` will be `http://www.example.com/foo.html` and `{primary_netloc}` will be `www.example.com`.

   When retriving requisites of `http://www.bar.org:8080/qux.html`, `{primary_url}` will be `http://www.bar.org:8080/qux.html` and `{primary_netloc}` will be `www.bar.org:8080`.

unignore

`!unignore IDENT PATTERN`, **`!unig IDENT PATTERN`** remove an ignore pattern:

```
> !unig 1q2qydhkeh3gfnrcxuf6py70b obnoxious\?foo=\d+
< Removed ignore pattern obnoxious\?foo=\d+ from job
  1q2qydhkeh3gfnrcxuf6py70b.
```

ignoreset

`!ignoreset IDENT NAME`, **`!igset IDENT NAME`** add a set of ignore patterns:

```
> !igset 1q2qydhkeh3gfnrcxuf6py70b blogs
< Added 17 ignore patterns to job 1q2qydhkeh3gfnrcxuf6py70b.
```

You may specify multiple ignore sets. Ignore sets that are unknown are, well, ignored:

```
> !igset 1q2qydhkeh3gfnrcxuf6py70b blogs, other
< Added 17 ignore patterns to job 1q2qydhkeh3gfnrcxuf6py70b.
< The following sets are unknown: other
```

Ignore set definitions can be found under db/ignore_patterns/.

ignorereports

`!ignorereports IDENT on|off`, **`!igrep IDENT on|off`** toggle ignore reports:

```
> !igrep 1q2qydhkeh3gfnrcxuf6py70b on
< Showing ignore pattern reports for job 1q2qydhkeh3gfnrcxuf6py70b.

> !igrep 1q2qydhkeh3gfnrcxuf6py70b off
< Suppressing ignore pattern reports for job
  1q2qydhkeh3gfnrcxuf6py70b.
```

Some jobs generate ignore patterns at high speed. For these jobs, turning off ignore pattern reports may improve both the usefulness of the dashboard job log and the speed of the job.

This command is aliased as `!igoff IDENT` and `!igon IDENT`. !igoff suppresses reports; !igon shows reports.

delay

`!delay IDENT MIN MAX`, **`!d IDENT MIN MAX`** set inter-request delay:

```
> !delay 1q2qydhkeh3gfnrcxuf6py70b 500 750
< Inter-request delay for job 1q2qydhkeh3gfnrcxuf6py70b set to [500,
  750 ms].
```

Delays may be any non-negative number, and are interpreted as milliseconds. The default inter-request delay range is [250, 375] ms.

concurrency

`!concurrency IDENT LEVEL`, **`!con IDENT LEVEL`** set concurrency level:

```
> !concurrency 1q2qydhkeh3gfnrcxuf6py70b 8
< Job 1q2qydhkeh3gfnrcxuf6py70b set to use 8 workers.
```

Adding additional workers may speed up grabs if the target site has capacity to spare, but it also puts additional pressure on the target. Use wisely.

yahoo

`!yahoo IDENT` set zero second delays, crank concurrency to 4:

```
> !yahoo 1q2qydhkeh3gfnrcxuf6py70b
< Inter-request delay for job 1q2qydhkeh3gfnrcxuf6py70b set to
  [0, 0] ms.
< Job 1q2qydhkeh3gfnrcxuf6py70b set to use 4 workers.
```

Only recommended for use when archiving data from hosts with gobs of bandwidth and processing power (e.g. Yahoo, Google, Amazon). Keep in mind that this is likely to trigger any rate limiters that the target may have.

expire

`!expire IDENT` for expiring jobs, expire a job immediately:

```
> !expire 1q2qydhkeh3gfnrcxuf6py70b
< Job 1q2qydhkeh3gfnrcxuf6py70b expired.
```

In rare cases, the 48 hour timeout enforced by ArchiveBot on archive jobs is too long. This command permits faster snapshotting. It should be used sparingly, and only ops are able to use it; abuse is very easy to spot.

If a job's expiry timer has not yet started, this command does not affect the given job:

```
> !expire 5sid4pgxkiu6zynhbt3q1gi2s
< Job 5sid4pgxkiu6zynhbt3q1gi2s does not yet have an expiry timer.
```

This is intended to prevent expiration of active jobs.

status

`!status` print job summary:

```
> !status
< Job status: 0 completed, 0 aborted, 0 in progress, 0 pending
```

`!status IDENT`, **`!status URL`** print information about a job or URL

For an unknown job:

```
> !status 1q2qydhkeh3gfnrcxuf6py70b
< Sorry, I don't know anything about job 1q2qydhkeh3gfnrcxuf6py70b.
```

For a URL that hasn't been archived:

```
> !status http://artscene.textfiles.com/litpacks/
< http://artscene.textfiles.com/litpacks/ has not been archived.
```

For a URL that hasn't been archived, but has children that have been processed before (either succesfully or unsuccessfully):

```
> !status http://artscene.textfiles.com/
< http://artscene.textfiles.com/ has not been archived.
< However, there have been 5 download attempts on child URLs.
< More info: http://www.example.com/#/prefixes/http://artscene.textfiles.com/
```

For an ident or URL that's in progress:

```
> !status 43z7a11vo6of3a7i173441dtc
<
< Downloaded 10.01 MB, 2 errors encountered
< More info at my dashboard: http://www.example.com
```

For an ident or URL that has been successfully archived within the past 48 hours:

```
> !status 43z7a11vo6of3a7i173441dtc
< Archived to http://www.example.com/site.warc.gz
< Eligible for rearchival in 30h 25m 07s
```

For an ident or URL identifying a job that was aborted:

```
> !status 43z7a11vo6of3a7i173441dtc
< Job aborted
< Eligible for rearchival in 00h 00m 45s
```

pending

!pending send pending queue in private message:

```
> !pending
< [privmsg] 2 pending jobs:
< [privmsg] 1. http://artscene.textfiles.com/litpacks/
              (43z7a11vo6of3a7i173441dtc)
< [privmsg] 2. http://example.blogspot.com/ncr
              (5sid4pgxkiu6zynhbt3q1gi2s)
```

Jobs are listed in the order that they'll be worked on. This command lists only the global queue; it doesn't yet show the status of any pipeline-specific queues.

whereis

!whereis IDENT, **!w IDENT** display which pipeline the given job is running on:

```
> !whereis 1q2qydhkeh3gfnrcxuf6py70b
< Job 1q2qydhkeh3gfnrcxuf6py70b is on pipeline
  "pipeline-foobar-1" (pipeline:abcdef1234567890).
```

For jobs not yet on a pipeline:

```
> !status 43z7a11vo6of3a7i173441dtc
< Job 43z7a11vo6of3a7i173441dtc is not on a pipeline.
```

CHAPTER 2

Indices and tables

- genindex
- modindex
- search