GNU Wget

presented by

Micah Cowan

november 15, 2010

About Micah Cowan

- Former maintainer of GNU Wget
- One-time co-maintainer of GNU Screen
- Current maintainer and author of GNU Teseq

micah@cowan.name
http://micah.cowan.name/

About GNU Wget

- Free software/software libre
- Fetches files off the web (HTTP(S) and FTP).
- Command line interface/batch-mode, non-interactive operation.
- Failure recovery
- Recursive downloads (website archival and such)

- Official page: http://www.gnu.org/software/wget/
- Wget Wgiki: http://wget.addictivecode.org/

My history with Wget

- Established bug-tracker usage
- Established the Wget Wgiki
  - 1.11 through 1.12(.0)

1.11

- More secure authentication
- Much more complete unit testing

1.12

- CSS parsing
- IRI support
- Better-delineated exit statuses

Restartable downloads

- Automatic retries
  - Only works if Wget knew the content-length ahead of time.
  - **--tries**/**-t**, for shaky connections, or to prevent too many retries.
  - Backs off on each retry an additional second, up to **--waitretry** (default 10) seconds.

Restartable downloads

- **--continue/-c**

  Caveats:

  - Can't work unless:
    - server gives file-size information, *and*
    - server supports continued downloads (ranged requests)
  - Some servers lie or contradict themselves about file size
  - **-c** can result in file corruption if file changed in the interim

Recursive-Descent Downloading

- **--recursive/-r**
- **--no-parent/-np** *(note: trailing slash in URL matters)*
- **--timestamping/-N**
- **--limit/-l**
- Fine-grained controls over which links to follow
  - **--page-requisites/-p**
  - Accept/Reject, Include/Exclude
  - Host controls (**-H**, **-D**)
- Link and filename conversion for local browsing (**-k** and **-E**)

Determining Save Paths

- recursive save location versus single-shot location
  - single-shot: won't overwrite pre-existing files by default, will add numeric suffix
  - forcing recursive-style with **-x**
  - forcing all-in-one-dir with **-nd**
- **-nH**
- **-P**
- **--cut-dirs**
- **-O**

Working Around Haters

- --user-agent/-U
- -e robots=off

Use Wget Conscientiously

- --limit-rate
- --wait, --random-wait

”Wput”-like Features

- **--post-data='***key=value&otherkey=nuthervalue***'**
- **--post-file=***filename*

  (where file contains *key=value&otherkey=nuthervalue*)
- No features whatsoever for uploading/deleting/etc via FTP
- No features for ”real” uploading, etc, via HTTP.

Cookies Support

- --save-cookies
- --load-cookies
- --keep-session-cookies

Debug mode

Good for:

- Viewing client and server headers
  - To see just the server headers, use **--server-response**/**-S**
- Debugging recursion problems:
  - Not the same host (use **-H**, with **-D**)
  - Rejected by rules (modify **-A**/**-R**/**-X**/**-I** settings)
  - Robots exclusions (**-e robots=off**)

Wget Config Files

- **~/.wgetrc** (override with WGETRC environment variable) and **/etc/wgetrc**
- Line by line, *key* = *value* syntax
- Useful for specifying persistent options, or just common batches of options (if you use WGETRC)
- Needs a **--config** option (thankfully, already have one in current dev sources).

Wget Shortcomings

“Fine-grained” link-following controls:

not fine-grained enough

- Always follows .htm/.html, ignoring **--accept/--reject**
  - Possibly deleting them afterward
  - This was in order to ensure we can find other links to our desired content: for instance, with **-A .pdf**, we'll traverse as many .html files as possible, saving all the PDF files we fine, and then delete the .htmls afterward.
    - *But:* it also downloads .html files at the edges of recursion limits, where you wouldn't follow any further links anyway.

Wget Shortcomings

“Fine-grained” link-following controls

- Always follows .htm/.html
  - What about when the accept/reject rules are intended to apply to the HTML files themselves?
  - What if the HTML-content files of a site aren't usually named .htm/.html? (examples: .php, .cfm, .asp, .jsp, *etc*)
- Can't apply accept/reject rules to query strings

  index.php?action=delete
  index.php?action=convert_to_pdf

  - Many wikis, etc, protect these links with ”nofollow”, but they shouldn't have to, and what about when we're not obeying robot exclusions?

Wget Shortcomings

“Fine-grained” link-following controls

- Needs a '**' wildcard in accept/rejects.

  Example: **--reject '**/DONTFOLLOW'**

- Wildcards aren't precise enough: need regex support.

Wget Shortcomings

- No HTTP/1.1 support *after 11 years as a standard!*
  - This is finally no longer the case in current dev sources (not released yet)
  - Prevents functioning with some sites; some CGI scripts and the like assume you support “chunked” transfer-encoding whether the client identified HTTP/1.1 support or not.

Wget Shortcomings

- No saved metadata, or stable-interface logging.
  - Makes it hard to write GUI wrappers around wget.
  - Would make mapping between URLs and local filenames easier (for instance, when **-nd** is used)
  - Remembered original timestamp and filesize, even after link conversion, or line-ending conversion
  - Remembered content types (AVI-as-HTML example)

Wget Shortcomings

- Not so Unixy after all?

  Wget is often praised as an example of the power and flexibility of the UNIX command line interface, but some of this reputation may not be deserved: the real power of UNIX's CLI lies in suites of small, composable tools with focused functionality

  - use an external link-parser
    - Support for emerging HTML standards
    - Parse links from arbitrary content-encodings
    - Light-weight JavaScript support?
  - filter links programatically—not just the accept/reject rules
  - Deal with arbitrary content-encodings, like gzip

Lessons Learned

- **-N** and **-O**
- **--auth-no-challenge**
- High visibility apparently != high developer activity. Who knew?
- Release breakage: 1.11 had broken restarts, despite plenty of “time” devoted to testing.
  - This motivated my drive to create much more complete unit testing

Issues Unique to GNU Projects

- Copyright assignments paperwork every time someone wants to submit a patch, or every time I move jobs
- Documentation: man page versus info

Questions?