wget Download and mirror

CONCEPTS

wget saves to a file by default and keeps going on its own: it retries, follows redirects and can recurse. curl prints to stdout and does not recurse; see the curl sheet for APIs.

Fetch URL retries, redirects
↓ saves
Local file -O, -P, -nd name it
↓ -r parses HTML
Follow links -l depth, -np, -A/-R
↓ -k rewrites
Offline copy -m -k -p

An existing file is not overwritten: wget saves as name.1, name.2. Use -c to resume, -N to refresh only newer files, or -nc to skip.

DOWNLOAD

`wget <url>`
Download to the current directory

`wget -O <file> <url>`
Save under a chosen name

`wget -P <dir> <url>`
Save into a directory

`wget -c <url>`
Resume an interrupted download

`wget -i <file>`
Read URLs from a file

`wget -q <url>`
Quiet, no output

`wget -nv <url>`
Short one-line output

`wget -b <url>`
Run in the background

`wget -O - <url>`
Write the body to stdout

`wget --show-progress -q <url>`
Quiet but keep the progress bar

`wget -nc <url>`
Skip if the file already exists

`wget -N <url>`
Only fetch if remote is newer

`wget --content-disposition <url>`
Use the server-suggested filename

RELIABILITY AND SPEED

`wget -t <n> <url>`
Number of retries (0 is unlimited)

`wget --retry-connrefused <url>`
Retry on refused connections

`wget --retry-on-http-error=503,429 <url>`
Retry on listed HTTP codes

`wget -T <sec> <url>`
Set all timeouts

`wget --waitretry=<sec> <url>`
Back off between retries

`wget --limit-rate=<rate> <url>`
Cap bandwidth (e.g. 500k, 2m)

`wget -w <sec> -i <file>`
Pause between downloads

RECURSIVE AND MIRROR

`wget -r <url>`
Follow links recursively (depth 5)

`wget -r -l <n> <url>`
Set recursion depth (0 or inf is unlimited)

`wget -np -r <url>`
Never ascend to the parent directory

`wget -m <url>`
Mirror: -r -N -l inf

`wget -k <url>`
Convert links for offline viewing

`wget -p <url>`
Also fetch images, CSS, scripts

`wget -nd -r <url>`
Flatten into one directory

`wget -nH --cut-dirs=<n> -r <url>`
Drop host and N path levels

`wget -A <list> -r <url>`
Accept only these extensions

`wget -R <list> -r <url>`
Reject these extensions

`wget --accept-regex=<re> -r <url>`
Accept URLs matching a regex

`wget --reject-regex=<re> -r <url>`
Skip URLs matching a regex

RECURSIVE AND MIRROR (cont.)

`wget -H -D <domains> -r <url>`
Span hosts but only these domains

`wget -E -k -p <url>`
Save one page fully for offline reading

HTTP, AUTH AND HEADERS

`wget --header='<Name: value>' <url>`
Add a request header

`wget --user=<u> --ask-password <url>`
Basic auth, prompt for password

`wget --http-user=<u> --http-password=<p> <url>`
HTTP-only credentials

`wget --load-cookies <file> <url>`
Send cookies from a file

`wget -U '<agent>' <url>`
Set the User-Agent

`wget --post-data='<k=v>' <url>`
Send a POST with form data

`wget -S <url>`
Show the server response headers

`wget --spider <url>`
Check a URL without saving it

TLS AND LOGGING

`wget --no-check-certificate <url>`
Skip certificate validation

`wget -o <file> <url>`
Write the log to a file

wget vs curl

Need	wget	curl
Default output	`file`	`stdout`
Resume	`-c`	`-C - -O`
Recurse	`-r`	`no`
API calls	`--method`	`-X -d`

RECIPES

Resume with retries

```
wget -c -t 0 --retry-connrefused
-T 30 <url>
```