

digital inspiration (/)

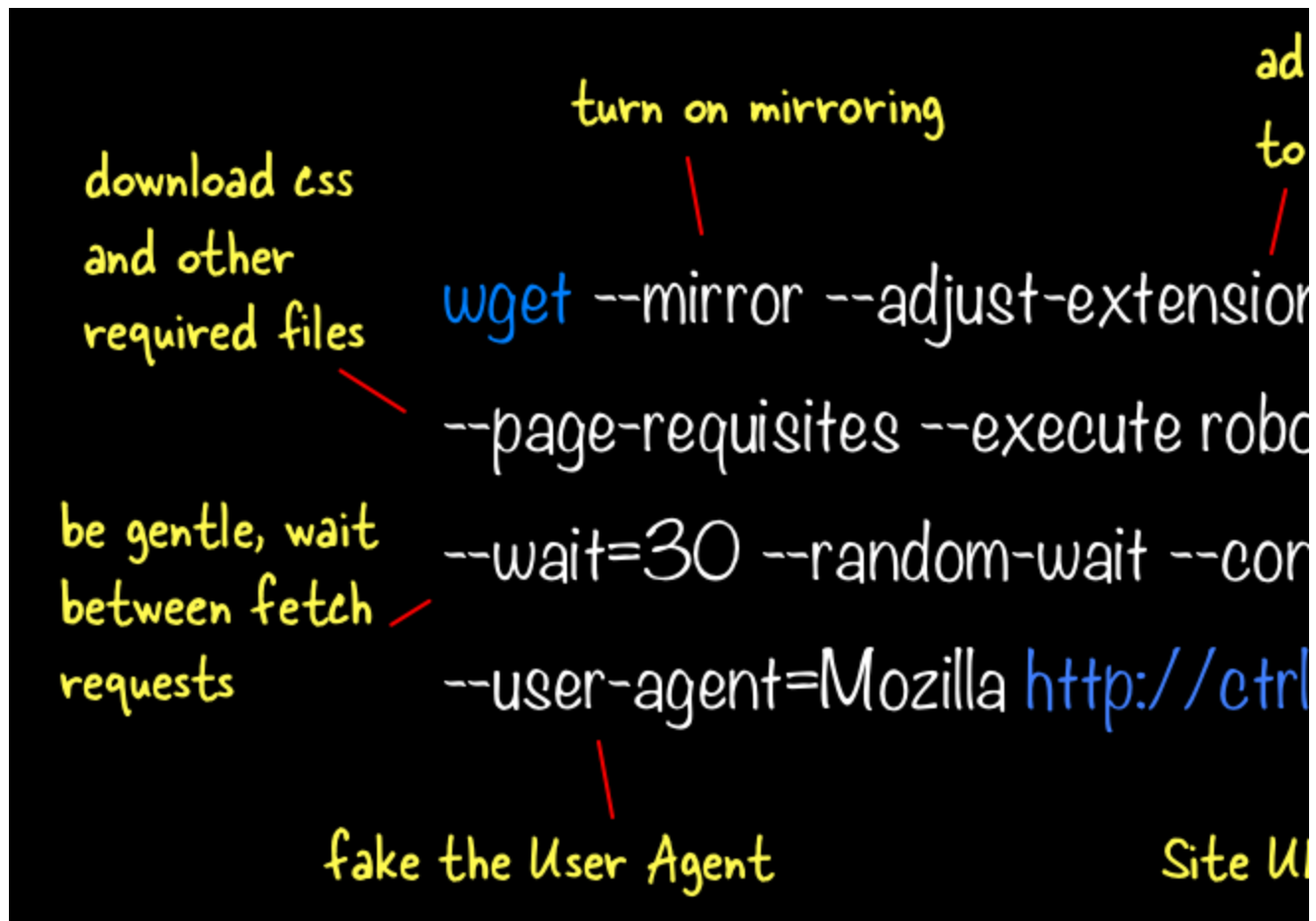
All the Wget Commands You Should Know

Wget lets you download Internet files or even mirror entire websites for offline viewing. Here are 20 practical examples for using the wget command.

DECEMBER 09, 2014

How do I download an entire website for offline viewing? How do I save all the MP3s from a website to a folder on my computer? How do I download files that are behind a login page? How do I build a mini-version of Google?

Wget (https://www.gnu.org/software/wget/manual/html_node/Overview.html#Overview) is a free utility – available for **Mac** (<http://brew.sh/>), **Windows** (<http://users.ugent.be/~bpuype/wget/>) and **Linux** (included) – that can help you accomplish all this and more. What makes it different from most download managers is that wget can follow the HTML links on a web page and recursively download the files. It is the **same tool** (<http://www.wired.com/2011/12/cables-scripts-manning/>) that a soldier had used to download thousands of secret documents from the US army's Intranet that were later published on the Wikileaks website.



Mirror an entire website with wget

Spider Websites with Wget – 20 Practical Examples

Wget is extremely powerful, but like with most other command line programs, the plethora of options it supports can be intimidating to new users. Thus what we have here are a collection of wget commands that you can use to accomplish common tasks from downloading single files to mirroring entire websites. It will help if you can read through the [wget manual \(https://img.labnol.org/di/wget.pdf\)](https://img.labnol.org/di/wget.pdf) but for the busy souls, these commands are ready to

execute.

1. Download a single file from the Internet

wget http://example.com/file.iso

2. Download a file but save it locally under a different name

wget --output-document=filename.html example.com

3. Download a file and save it in a specific folder

wget --directory-prefix=folder/subfolder example.com

4. Resume an interrupted download previously started by wget itself

wget --continue example.com/big.file.iso

5. Download a file but only if the version on server is newer than your local copy

wget --continue --timestamping wordpress.org/latest.zip

6. Download multiple URLs with wget. Put the list of URLs in another text file on separate lines and pass it to wget.

```
wget --input list-of-file-urls.txt
```

7. Download a list of sequentially numbered files from a server

```
wget http://example.com/images/{1..20}.jpg
```

8. Download a web page with all assets – like stylesheets and inline images – that are required to properly display the web page offline.

```
wget --page-requisites --span-hosts --convert-links --adjust-extension http://example.com/dir/file
```

Mirror websites with Wget

9. Download an entire website including all the linked pages and files

```
wget --execute-robots=off --recursive --no-parent --continue
```

--no-clobber http://example.com/

10. Download all the MP3 files from a sub directory

*wget --level=1 --recursive --no-parent --accept mp3,MP3
http://example.com/mp3/*

11. Download all images from a website in a common folder

*wget --directory-prefix=files/pictures --no-directories --recursive
--no-clobber --accept jpg,gif,png,jpeg http://example.com/images/*

12. Download the PDF documents from a website through recursion but stay within specific domains.

*wget --mirror --domains=abc.com,files.abc.com,docs.abc.com
--accept=pdf http://abc.com/*

13. Download all files from a website but exclude a few directories.

```
wget --recursive --no-clobber --no-parent --exclude-directories  
/forums,/support http://example.com
```

Wget for Downloading Restricted Content

Wget can be used for downloading content from sites that are behind a login screen or ones that check for the HTTP referer and the User Agent strings of the bot to prevent screen scraping.

14. Download files from websites that check the User Agent and the HTTP Referer

```
wget --refer=http://google.com --user-agent="Mozilla/5.0  
Firefox/4.0.1" http://nytimes.com
```

15. Download files from a password protected (<https://ctrlq.org/code/19247-password-protect-wordpress-admin>) sites

```
wget --http-user=labnol --http-password=hello123  
http://example.com/secret/file.zip
```

16. Fetch pages that are behind a login page. You need to replace user and password with the actual form fields while the URL should point to the Form Submit (action) page.

```
wget --cookies=on --save-cookies cookies.txt --keep-session-cookies  
--post-data 'user=labnol&password=123' http://example.com  
/login.php  
wget --cookies=on --load-cookies cookies.txt --keep-session-cookies  
http://example.com/paywall
```

Retrieve File Details with wget

17. Find the size of a file without downloading it (look for Content Length in the response, the size is in bytes)

```
wget --spider --server-response http://example.com/file.iso
```

18. Download a file and display the content on screen without saving it locally.

```
wget --output-document - --quiet google.com/humans.txt
```

19. Know the last modified date of a web page (check the Last Modified tag in the HTTP header).

```
wget --server-response --spider http://www.labnol.org/
```

20. Check the links on your website to ensure that they are working. The spider option will not save the pages locally.

```
wget --output-file=logfile.txt --recursive --spider http://example.com
```

Also see: **Essential Linux Commands** (<https://www.labnol.org/software/linux-commands/19028/>)

Wget – How to be nice to the server?

The wget tool is essentially a spider that scrapes / leeches web pages but some web hosts may block these spiders with the robots.txt files. Also, wget will not follow links on web pages that use the **rel=nofollow** (<https://www.labnol.org/internet/drop-nofollow-from-internal-links/14107/>) attribute.

You can however force wget to ignore the robots.txt and the nofollow directives by adding the switch **--execute robots=off** to all your wget commands. If a web host is blocking wget requests by looking at the User Agent string, you can always fake that with the **--user-agent=Mozilla** switch.

The wget command will put additional strain on the site's server because it will continuously traverse the links and download files. A good scraper would therefore limit the retrieval rate and also include a wait period between consecutive fetch requests to reduce the server load.

```
wget --limit-rate=20k --wait=60 --random-wait --mirror example.com
```

In the above example, we have limited the download bandwidth rate to 20 KB/s and the wget utility will wait anywhere between 30s and 90 seconds before retrieving the next resource.



Finally, a little quiz. What do you think this wget command will do?

wget --span-hosts --level=inf --recursive dmoz.org

COMMENT ([HTTPS://DOCS.GOOGLE.COM/FORMS/D/E/1FAIPQLSDXTPLMAOPVRUZKGE_XV7LZXK6XUFBEE5](https://docs.google.com/forms/d/e/1FAIPQLSDXTPLMAOPVRUZKGE_XV7LZXK6XUFBEE5))

NEWSLETTER (/NEWSLETTER)

HOME (/) / TECH GUIDES (/TAG/GUIDE/) / LINUX ([HTTPS://WWW.LABNOL.ORG/TAG/LINUX/](https://www.labnol.org/tag/linux/))

 ([HTTPS://TWITTER.COM/INTENT/TWEET?SOURCE=HTTPS%3A%2F%2FWWW.LABNOL.ORG%2FSOFTWARE%2FWGET-COMMAND-EXAMPLES%2F28750%2F&VIA=LABNOL&RELATED=LABNOL&URL=HTTPS%3A%2F%2FWP.ME%2FP4F8F-7TI&TEXT=ALL+THE+WGET+COMMANDS+YOU+SHOULD+KNOW](https://twitter.com/intent/tweet?source=https%3A%2F%2Fwww.labnol.org%2Fsoftware%2Fwget-command-examples%2F28750%2F&via=labnol&related=labnol&url=https%3A%2F%2Fwp.me%2FP4F8F-7TI&text=all+the+wget+commands+you+should+know))  ([HTTP://WWW.FACEBOOK.COM/SHARER.PHP?U=HTTPS%3A%2F%2FWWW.LABNOL.ORG%2FSOFTWARE%2FWGET-COMMAND-EXAMPLES%2F28750%2F](http://www.facebook.com/sharer.php?u=https%3A%2F%2Fwww.labnol.org%2Fsoftware%2Fwget-command-examples%2F28750%2F))



Amit Agarwal (<https://www.labnol.org/about/>) is a web geek (<https://ctrlq.org/>), ex-columnist for The Wall Street Journal and founder of Digital Inspiration (<https://www.labnol.org/>), a hugely popular tech how-to website since 2004. He holds an engineering degree in Computer Science from IIT and happens to be the first professional blogger

(<https://yourstory.com/2015/07/techie-tues-amit-agarwal-labnol/>) in India. He's been p Lifehacker (<http://lifehacker.com/im-amit-and-this-is-how-i-work-1506511234>). With experience in software development, / authored several popular Google (<https://www.labnol.org/internet/best-google-docs-add-ons/28440/>) that are deployed in

the biggest companies and universities worldwide with over 250,000 installations. Download the PDF

brochure (<http://go.ctrlq.org/GoogleAutom> know more.

Email: amit@labnol.org



(<https://twitter.com/labnol>)



RECOMMENDED

RELATED

WHAT'S NEW

EVERGREEN

GOOGLE ADDONS

Moving Your Website from Google Pages to Google Sites

(<https://www.labnol.org/internet/move-from-google-pages-to-sites/82>

2009-04-17

Web Scraping Reddit with Google Scripts (<https://www.labnol.org/internet/web-scraping-reddit/28369/>)

2014-01-09

How to Download PostSecret Archives (<https://www.labnol.org/internet/postsecret-archives/11963/>)
Install Google Web Fonts on your Computer (<https://www.labnol.org/software/google-fonts-on-computer/19780/>)

2012-07-15

2012-07-22

Create a Tree View of your Google Drive (<https://www.labnol.org/internet/google-drive-tree/21198/>)

2014-11-25

SnagIt Tips and Tricks - Capture Great Looking Screenshots & more
(<https://www.labnol.org/software/tutorials/snagit-tips-tricks-screen-capture/1936/>)

2007-12-14

We build bespoke solutions that use the capabilities and the features of Google Apps for automating business processes and driving business productivity.

GET IN TOUCH (HTT



HOME	CONTACT US	Connect with us
(/)	(https://docs.google.com/forms	(https://twitter.com/lab
ABOUT	/d/e/1FAIpQLSdxtplmAopvRUZKGe_xV7IZXK6XufBee98xNMrrKH0J	(https://www.faceboo
(/about/)	/viewform?entry.1823440910&entry.910693880&entry.1551994540=	/digital.inspiration)
SETUP	entry.1031080712=https:	(http://feeds.labnol.o
(/setup/)	//www.labnol.org/software/wget-	/labnol)
F.A.Q.	command-examples/28750	(https://www.youtube
(/faq/)	/%23footer&entry.857400982&	/labnol)
	entry.467659801&	
	entry.1471550472&	
	entry.2073450133&	
	entry.1252392100&	
	entry.2066640595&	
	entry.893931147)	
	NEWSLETTER	
	(/newsletter)	

GOOGLE AUTOMATION

(<https://ctrlq.org/>)

TOP LISTS

(<https://digitalinspiration.com/>)

© 2004-2017 Digital Inspiration, tech à la carte. Made in India.