Tutorials

Download entire websites easy

[GNU Wget](http://www.gnu.org/software/wget/) is a nice tool for downloading resources from the internet. The basic usage is `wget url`:

> `wget http://linuxreviews.org/`

The power of [wget](http://community.linuxmint.com/software/view/wget) is that you may download sites recursive, meaning you also get all pages (and images and other data) linked on the front page:

> `wget -r http://linuxreviews.org/`

But many sites do not want you to download their entire site. To prevent this, they check how browsers identify. Many sites refuse you to connect or send a blank page if they detect you are not using a web-browser. You might get a message like:

> *Sorry, but the download manager you are using to view this site is not supported. We do not support use of such download managers as flashget, go!zilla, or getright*

There is a very handy `-U` option for sites like this. Use

> `-U My-browser`

to tell the site you are using some commonly accepted browser:

> wget -r -p -U Mozilla http://www.stupidsite.com/restricedplace.html

A web-site owner will probably get upset if you attempt to download his entire site using a simple

> `wget http://foo.bar`

command. However, the web-site owner will not even notice you if you limit the download transfer rate and pause between fetching files.

To make sure you are not manually added to a blacklist, **the most important command line options** are` –limit-rate=` and` –wait= .`

To pause 20 seconds between retrievals you should add

> `–wait=20`

and to limit the download rate use something like

> `–limit-rate=20K`

as this option defaults to bytes, add K to set KB/s.

Example:

> `wget –wait=20 –limit-rate=20K -r -p -U Mozilla http://www.stupidsite.com/restricedplace.html`

A very handy option that guarantees wget will not download anything from the folders beneath the folder you want to acquire is:

> `–no-parent`

Use this to make sure wget does not fetch more than it needs to if you just want to download the files in a folder.

Read the [manual page](http://linuxreviews.org/man/wget/) for wget to learn more about GNU Wget. The full official manual is available [here](http://www.gnu.org/software/wget/manual/).

The original version of this how-to is available at http://linuxreviews.org/quicktips/wget/wget.en.pdf

Copyright © 2000-2004 [Øyvind Sæther](http://oyvinds.everdot.org/). Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled [“GNU Free Documentation

License”](http://www.gnu.org/licenses/fdl.html).

From:
https://wiki.tromjaro.alexio.tf/ - **TROMjaro wiki**

Permanent link:
https://wiki.tromjaro.alexio.tf/doku.php?id=tutorials&rev=1638617648

Last update: **2021/12/04 11:34**