Published September 2021
at Iffy Books

Version 0.9

Download this zine as a PDF here:
https://iffybooks.net/wget

No rights reserved.

PUBLIC DOMAIN

# Download a website with Wget

*a do-it-yourself guide*
*from Iffy Books*

IFFY BOOKS

*319 N. 11th St. 3D*
*Philadelphia, PA 19107*

*Join our email list at*
*iffybooks.net*

*Follow @iffybooks*
*on social media*

*Send corrections to*
*iffybooks@protonmail.com*

And here's a Wget command that uses it:

```
wget -r -np -e robots=off --wait=0.3
--random-wait --user-agent="Mozilla/5.0
(Macintosh; Intel Mac OS X 10.15; rv:78.0)
Gecko/20100101 Firefox/78.0" https://
iffybooks.net/images/
```

(This page doesn't use backslashes for line endings because the user-agent string is so long!)

## ✿ Put it all together ✿

When you combine the options you've learned, your final Wget command for mirroring a website might look something like this:

```
wget -c -r -l inf --convert-links -np
--page-requisites -erobots=off --wait=0.82
--random-wait --user-agent="Mozilla/5.0
(Macintosh; Intel Mac OS X 10.15; rv:78.0)
Gecko/20100101 Firefox/78.0" "https://
iffybooks.net/images/"
```

You can also use the `--mirror` or `-m` option, which would replace `-r` and `-l inf` in the command above.

Tip: You might want to assemble your Wget command in a text editor first, then copy and paste it into the terminal.

## ✿ Read the Wget docs ✿

To view the help page for Wget, enter the following command:

```
wget -h
```

Try running the following command to view the full Wget manual in the terminal:

```
man get
```

Or find the Wget manual at the following URL:

https://www.gnu.org/software/wget/manual/wget.html

## ★ Ignore the robots.txt file ★

**`--execute robots=off`** or
**`-e robots=off`**

Sometimes when you try to mirror a website, Wget will immediately stop running and won't download anything. Often that's because the website has a **robots.txt** file that asks web crawlers to stay away. By default, Wget will obey these rules.

To ignore a website's **robots.txt** file, you can use the option **`--execute robots=off`** or **`-e robots=off`**. The following command will download files from a single directory while ignoring the **robots.txt** file.

```
wget -r --no-parent -e robots=off \
    https://iffybooks.net/images/
```

## ★ Spoof your user agent ★

**`--user-agent="`*`agent-string`*`"`** or
**`-U "`*`agent-string`*`"`**

Replace ***`agent-string`*** with a user-agent string.

Every time you request a file over the internet, your computer generates an HTTP request. And every HTTP request contains a User-Agent header field that specifies the application and operating system being used. Some websites block requests that don't come from a program like Wget, in which case you'll need to switch up your user-agent string.

You can go to **https://whatsmyua.info** to see your own browser's user-agent string. Here's an example:

```
Mozilla/5.0 (Macintosh; Intel Mac OS X 10.15;
   rv:78.0) Gecko/20100101 Firefox/78.0
```

# Wget is a command-line tool for downloading files from the internet.

If you run a website, wouldn't it be nice to have a backup just in case? Wget can help.

Maybe you've downloaded a collection of PDFs or MP3 files from the web by right-clicking 100 links in a row. With Wget, you can grab them all with one line of code.

You can use Wget to download one file, or a thousand, or a billion. It's super powerful and reliable, even if you're using a slow internet connection. And it's easy to get started!

Downloading a single file with Wget looks like this: **`wget`**, then a space, then a URL.

```
wget https://iffybooks.net/Wget-zine.pdf
```

If you add a few options to the command, Wget can automatically mirror an entire website.

```
wget --mirror --page-requisites --wait=0.5 \
     --random-wait https://iffybooks.net
```

This backslash means the command continues on the next line.

Wget is part of the GNU Project, and it's been around since 1996. Wget is widely used by programmers, as well as by journalists, archivists, and academics. It's also been used by activists, including Chelsea Manning and Aaron Swartz.

This zine will guide you through your first few downloads with Wget, even if you've never used the command line before. Good luck!

## ✱ Install Wget on macOS ✱

1

First you'll need to install the Homebrew package manager. If you already have Homebrew, skip to step 4.

Go to **https://brew.sh** and read a bit about Homebrew. When you're ready to install, highlight the line of code under "Install Homebrew" and copy it. Here's what it looks like:

```
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
```

2

Open a new window in **Terminal**. If you haven't used Terminal recently, you can find it in the directory **/Applications/Utilities/**. Or press command+space and search for "Terminal."

Paste the line of code from the previous step into the terminal window and press enter. You'll need to enter your password.

3

You may be asked to install "XCode Command Line Tools" (~3GB). Follow the prompts to install it. Then follow the prompts to finish installing Homebrew. This will take a while, at least 5 minutes.

4

To install wget using Homebrew, type the following command in your terminal window and press enter.

```
brew install wget
```

To address this issue, you can use the **--adjust-extension** or **-E** option to give every file an explicit file extension. (Delete your previous site mirror before running the following command.)

```
wget -m --continue --convert-links \
--adjust-extension https://iffybooks.net
```

## ✱ Follow links to all other hosts ✱

**--span-hosts** or **-H**

By default, Wget will ignore links to domains other than the one in the URL you're using. The command below downloads links to all other hosts, using a recursion depth of 1.

```
wget -r --level=1 --continue --wait=0.2 \
    --span-hosts https://iffybooks.net
```

Be careful! You typically won't use this option without the --**domains** option (below), because Wget will end up trying to download the entire internet.

## ✱ Follow links to specific hosts ✱

**--domains=*domain-list*** or **-d *domain-list***

When you're using the **--span-hosts** option, you can add the **--domains** or **-d** option to provide a list of additional domains that Wget is allowed to crawl. This feature comes in handy for sites that host images and media files externally.

```
wget -r --level=3 --continue --wait=0.2 \
--span-hosts --domains="principalhand.org" \
            https://iffybooks.net
```

To allow links to multiple hosts, you can use several domains separated by commas.

## ★ Download supporting files ★

### `--page-requisites` or `-p`

By default, Wget will only download files that are linked directly from a website. To download CSS files and other resources that help render pages, use the `--page-requisites` or `-p` option.

```
wget -r -np --wait=3 --random-wait \
--page-requisites https://iffybooks.net/images/
```

## ★ Reconnect links ★

### `--convert-links` or `-k`

Once you've downloaded a local mirror of a website, open any HTML file in a browser and try to navigate around the site. If you find absolute links that lead back to the original site, you may want to use the `--convert-links` or `-k` option to replace those absolute links in your mirror with relative ones.

```
wget -m --continue --convert-links https://iffybooks.net
```

The command above uses `--continue`, which means it won't download any files that have already been downloaded.

## ★ Adjust file extensions ★

### `--adjust-extension` or `-E`

In many cases, files downloaded with Wget will end up without the file extensions you'd normally expect. For example, an HTML file might have a name like 'index.html?p=28', including a query string after the original filename. The same thing can happen with CSS files.

5

When installation is finished, type the following command in the terminal and press enter. If Wget is installed, you'll see a help message with a list of options.

```
wget -h
```

## ★ Install Wget on Windows ★

1

Go to the following URL and download the most recent version of Wget. You'll want the 64-bit version if you're using a newer computer.

https://eternallybored.org/misc/wget/

Why is the most up-to-date version of an important program hosted on someone's personal website? Welcome to the world of Windows free software binaries. If you're curious, you can find the URL above linked from the official GNU Wget site.

https://www.gnu.org/software/wget/

2

Next you'll move the file **wget.exe** to the directory **C:\Windows\System32** to make it executable from the command prompt.

To navigate to your System32 directory, open a File Explorer window and double click the name of the current directory at the top of the window. Then type **C:\Windows\System32** and press enter.

Find **wget.exe** in your Downloads directory and drag it into **C:\Windows\System32**. You'll need to provide administrator permission to move the file.

3

Open the Windows search box and type “cmd.” Click **Command Prompt** to open a DOS terminal window, type the following command, and press enter. If Wget is installed, you'll see a help message with a list of options.

```
wget -h
```

## Install Wget on Linux

If you're using Linux, Wget is probably already installed. Enter the following command in a terminal window and you should see the Wget help message.

```
wget -h
```

If you're using a stripped-down copy of Linux that doesn't have Wget, you can install it with your package manager. On a Debian-based OS like Ubuntu, that looks like this.

```
sudo apt-get install wget
```

## Create a new directory

To keep things neat and tidy, start by creating a new directory on your desktop. If you're using macOS or Linux, use the following **cd** command to change your current directory to the desktop.

```
cd ~/Desktop
```

If you're using Windows, enter the following in the command prompt instead.

```
cd Desktop
```

Type the following **mkdir** command and press enter to make a new directory called **Wget_Downloads**.

```
mkdir Wget_Downloads
```

downloads. In other cases, hitting a server with requests too quickly can make the site go down temporarily.

## Wait a random duration

**`--random-wait`**

Waiting the same amount of time between downloads might make it obvious that you're using an automated program. If you add the `--random-wait` option, Wget will wait for a random interval after each download, from 0.5 to 1.5 times the number of seconds specified with the `--wait` option.

```
wget -r -np --wait=3 --random-wait \
https://iffybooks.net/images/
```

## Mirror a site with `--mirror`

**`--mirror`** or **`-m`**

The `--mirror` or `-m` option is a handy shortcut, combining several options that work well for mirroring a website. Most importantly, it turns on recursion and sets an infinite recursion depth.

The command below will create a local mirror of iffybooks.net, waiting 0.2 seconds after each download.

```
wget --mirror --wait=0.2 https://iffybooks.net
```

If you're just creating a backup, the command above might be all you need. If you want to view your local mirror in a browser or host it somewhere else, consider using the next few options as well.

## ✭ Download files in a specific directory ✭

**`--no-parent`** or **`-np`**

You can use the **`--no-parent`** or **`-np`** option along with **`--recursive`** or **`-r`** to have Wget only download files that fall under the directory specified in the URL. In other words, Wget won't follow links to any parent directories.

The following command downloads everything under the directory /images/ on iffybooks.net.

```
wget -r --no-parent https://iffybooks.net/images/
```

Here's the short version:

```
wget -r -np https://iffybooks.net/images/
```

## ✭ Wait between downloads ✭

**`--wait=`*`seconds`*** or **`-w`** ***`seconds`***

When you're downloading lots of files from the same site, it's a good idea to have Wget pause between downloads. The command below will wait for 3 seconds after each download.

```
wget -r -np --wait=3 https://iffybooks.net/images/
```

This one, using the short version of the option, pauses for 0.2 seconds after each download.

```
wget -r -np -w 0.2 https://iffybooks.net/images/
```

You should use the **`--wait`** option every time you download a website recursively, as a courtesy to the person running the site. Some sites will block you if you download files too quickly, in which case you'll *need* to wait between

Now use **`cd`** to change your current directory to the one you just created.

```
cd Wget_Downloads
```

Go to the desktop in the Finder/File Explorer and you should see your new directory.

## ✭ Download a single file ✭

To download a file from the web, type **`wget`** followed by a space, then the file's URL. The file will be downloaded to the current directory.

```
wget https://iffybooks.net/Wget-zine.pdf
```

If you use a web page URL that doesn't include a filename, Wget will download the page's source code to an HTML file called **index.html**.

```
wget https://iffybooks.net
```

Run the command above, then go to the **Wget_Downloads** directory on your desktop and you'll see a file called **index.html**. Open the file in a text editor and skim through the source code. If you open it in a web browser, you'll see a version of the page with odd formatting and missing images.

Go to your terminal window and press the **up arrow** to see your previous command. Press enter to run it again. Wget will download the same file with the name **index.html.1** instead of overwriting the original file.

If the URL you're using contains so-called escape characters, such as question marks (?), ampersands (&), or spaces, you'll need to put quotation marks around it.

```
wget "https://iffybooks.net/File with spaces.jpg"
```

## ✿ Add options to your Wget command ✿

A Wget command always begins with **wget** and ends with a URL. To modify Wget's behavior, you can add options in between, separated by spaces.

For example, the **--debug** option tells Wget to output more detailed information about each download. Run the command below and see what happens.

```
wget --debug https://iffybooks.net/Wget-zine.pdf
```

Most Wget options have a short form and a long form, which can be used interchangeably. The long version of an option begins with two hyphens, like **--debug** in the example above. The short version uses a single hyphen, like **-d** in the command below.

```
wget -d https://iffybooks.net/Wget-zine.pdf
```

You can add as many options to your command as you need.

## ✿ Continue an incomplete download ✿

**--continue** or **-c**

If your download gets interrupted partway through, use the **--continue** or **-c** option to continue the download where you left off.

```
wget --continue https://iffybooks.net/Wget-zine.pdf
```

Here's the short version:

```
wget -c https://iffybooks.net/Wget-zine.pdf
```

## ✿ Download a website recursively ✿

**--recursive** or **-r**

In addition to downloading individual files, Wget can work like a web crawler, navigating through a site recursively and downloading everything it finds.

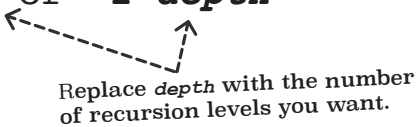
```
wget --recursive https://iffybooks.net
```

By default, Wget uses a recursion depth of 5. That means Wget will visit pages 5 directories deep from the URL you provide.

As a general rule, you shouldn't use the **--recursive** or **-r** option without using the **--wait** option (introduced below) to pause between downloads. Otherwise you may send requests too quickly and overload the server.

## ✿ Set the recursion depth ✿

**--level=*depth*** or **-l *depth***




To adjust the recursion depth, you can use the **--level** or **-l** option and specify the number of directory levels Wget will explore. The following command will download files 3 directories deep.

```
wget --recursive --level=3 https://iffybooks.net
```

Here's the short version of the same command:

```
wget -r -l 3 https://iffybooks.net
```

Note that the long form of the option above uses an equals sign to set the depth value (**--level=3**), while the short form uses a space instead (**-l 3**).

To use an infinite recursion depth, set the **--level** value to 0 (zero) or "inf".

```
wget -r --level="inf" https://iffybooks.net
```