Make Offline Mirror of a Site using `wget`

Sometimes you want to create an offline copy of a site that you can take and view even without internet access. Using `wget` you can make such copy easily:

```
wget --mirror --convert-links --adjust-extension --page-requisites
--no-parent http://example.org
```

Explanation of the various flags:

- `--mirror` – Makes (among other things) the download recursive.
- `--convert-links` – convert all the links (also to stuff like CSS stylesheets) to relative, so it will be suitable for offline viewing.
- `--adjust-extension` – Adds suitable extensions to filenames (`html` or `css`) depending on their content-type.
- `--page-requisites` – Download things like CSS style-sheets and images required to properly display the page offline.
- `--no-parent` – When recursing do not ascend to the parent directory. It useful for restricting the download to only a portion of the site.

Alternatively, the command above may be shortened:

```
wget -mkEpnp http://example.org
```

Note: that the last `p` is part of `np` (`--no-parent`) and hence you see `p` twice in the flags.

SHARE THIS:

Twitter LinkedIn Facebook Email More

Installing OrderNet Pro on Debian Jessie
July 23, 2016
In "Tutorials"

Downloading Audio/Songs from YouTube
August 16, 2013
In "Tips"

Tracking MediaWiki External Links Statistics using Google Analytics
July 16, 2007
In "Google Analytics"

May 2, 2014 Guy Tips wget

27 thoughts on “Make Offline Mirror of a Site using `wget`”

David Wolski
July 7, 2014 at 13:59

wget usually doesn’t work very well for complete offline mirrors of website. Due to its parser there is always somethings missing, i.e. stylesheets, scripts, images. It simply isn’t the right tool for this task.
HTTrack is much slower than wget but a powerful parser. It’s GPL and available in most Linux-Distributions.
Documentation and sorce-code is available at http://www.httrack.com

Pingback: Make an Offline Mirror of a Site Using `wget` - John Haynes

Pingback: Linux 如何抓取網頁頁面 並 將相關連結置換 或 完整抓取下來 - Tsung's Blog

Pingback: Mirror | stigmatedbrain's corner

Pingback: Download a complete single page with wget - justnorris

bhl
September 11, 2015 at 20:12

I second David Wolski’s comment. HTTrack is an outstanding website mirroring tool. I like it because it performs incremental updates. Nothing like sucking down the Wash-

ington Post without adverts.

Pingback: web archiving resources for NDSA NE crew (and anyone else reading this!) | Archive Hour

---

Pingback: Niente stronzate ©



Imam Suharjo

May 17, 2016 at 05:01

Thanks You for helping us 🙂
It's wget -mkEpnp in my server

---

mina

October 24, 2016 at 03:37

thankss

Herbert Prollermann

December 20, 2016 at 19:33

Wget is a great tool, very helpfull to make some website backups for my private archive.

---

codywohlers

January 3, 2017 at 15:42

Thanks for the article!

You can also –wait and –random-wait to reduce the load on the server.

Mauro

January 11, 2017 at 18:47

Thank you so much Guy!

Pingback: Download a complete single page with wget

JGM

July 28, 2017 at 12:18

Thanks!

Pingback: 用wget抓取整站 | my life, my blog

Juan

February 18, 2018 at 15:15

Thank you very much!

---

xmetalx

March 12, 2018 at 21:14

Si non il y a un script RipTool sur https://www.opendesktop.org/p/1218850/

Ashutosh

November 10, 2018 at 08:39

how to use this for https websites ?

Ali

December 4, 2018 at 09:41

@Ashutosh
For https website, just add parameter –no-check-certificate

Example:
wget -mkEpnp –no-check-certificate https://example.com

Gwyneth Llewelyn

July 1, 2019 at 22:33

I saw the comment related to HTTrack only after reading this very useful article (and successfully copying 99% of a website written in ColdFusion, the remaining 1% being embedded JavaScript which had to be done manually; also, moving everything to HTTPS took me a minute or so!).

Unfortunately, HTTrack made sense in 2014 (when this article was written), but it stopped being developed in 2017 (last commit on github) and has 112 pending issues (a bad sign — it's probably abandoned by now). One major issue with HTTrack is the apparent lack of support of HTML5 (or at least incomplete support for the new tags).

wget continues to be thoroughly developed, and, although I haven't tried it personally (I'm mostly copying 'legacy' websites…), it seems to be able to deal with HTML5 tags so far as one 'forces' wget to identify itself as a recent version of, say, Chrome or Firefox; if it identifies itself by default, the webserver it connects too may simply think that it's a very old browser trying to access the site and 'simplify' the HTML being passed back (i.e. 'downgrading' it to HTML4 or so). This, of course, is not an issue with wget *per se* but rather the way webservers (and web designers!) are getting more and more clever in dealing with a vast variety of users, browsers, and platforms.

Also, contemporary versions of wget (which means mid-2019 by the time I'm writing this comment!) will have no trouble 'digging deep' to extract JS and CSS files etc. Obviously it cannot make miracles and doesn't deal with everything; I had some issues with imagemaps, for instance (something nobody uses these days), as well as HTML generated on-the-go by Javascript. And of course there is a limit to what it can actually do with very complex and dynamic websites which adjust their content to whatever browser the user has, page by page — especially on those cases where the different versions of the same page all have the same URL (a bad practice IMHO). Still, it remains useful for a lot of situations, and the results are better than what you get out of archive.org…

I just wanted to point this out since this article is old but still relevant for today's wget — and sadly HTTrack was abandoned and isn't an option any longer…

Pingback: Mirroring a site using Wget – The Publishing Project

BobDodds

December 20, 2019 at 09:11

I tried pavuk, which can handle javascript but it got confused so I went back to wget. Somebody may want to further research pavuk with js.

I tried wget with different parameters, saw a lot of errors. Your setup, -mkEpnp. is now downloading smoothly.
alias wgetMirror="/usr/bin/wget -o wget.log -mkEpnp –wait=9 –user-agent='Mozilla/5.0 (compatible; Googlebot/2.1; +http://www. google.com/bot.html)' –no-check-certificate"

I'm trying to save somebody's wordpress site which he lost control of, so we cannot ssh into vps and tar gzip. I have wordpress on another vps for him.

Amba

July 15, 2020 at 16:01

After many unsuccessful parameter combination attempts,
your command worked perfectly on my system!
Scientific Linux release 6.10 (Carbon)
GNU Wget 1.12

Pingback: Exporting from CSV format into Markdown – the long way.

---

anonymous user

December 21, 2020 at 10:10

sometimes an additional `--compression=auto` is required to handle gzip. otherwise, you'd get a single `index.html.gz`

Christer

February 17, 2021 at 11:56

I just created a static site with some 900 pages from a Ruby on Rails application using wget with the proposed parameters. Worked perfectly! The two things I really like are that it managed to transform internal links to relative (adding "../" or "../../" as needed)

and that extensions were added to the file names.
This only took a few minutes (on OS X Big Sur) with my Rails application running in development mode on the same machine.

This site uses Akismet to reduce spam. Learn how your comment data is processed.
