# How to crawl website with Linux wget command

## What is wget

Wget is a free utility for non-interactive download of files from the Web.It supports HTTP, HTTPS, and FTP protocols, as well as retrieval through HTTP proxies.

## Download a web page

```
wget http://dumps.wikimedia.org/dewiki/20140528/
```

The result will be saved in a file "*.html" in the current directory

## Download full web site

```
mkdir -p /data/sample
```

```
cd /data/sample
```

```
wget --no-clobber --convert-links --random-wait -r -p --level 1 -E -e robots=off -U r
```

### Option

-b: runs it in background and cant see progress

-c: continue getting a partially-downloaded file.  This is useful when you want to finish up a download started by a previous instance of Wget, or by another program

-e: robots=off: act like we are not a robot - not like a crawler - websites dont like robots/crawlers unless they are google/or other famous search engine

-E: gets the right extension of the file, without most html and other files have no extension
-p: get all the page requisites. e.g. get all the image/css/js files linked from the page.
-r: ecursive - downloads full website
-U: pretends to be just like a browser Mozilla is looking at a page instead of a crawler like wget

-nd: do not create a hierarchy of directories when retrieving recursively. With this option turned on, all files will get saved to the current directory, without clobbering
-np: wget will not follow links up the url. e.g. it will not follow a link from devopsa.net/linux/curl.html to devopsa.net/linux.html

--connect-timeout: Set the connect timeout to seconds seconds.  TCP connections that take longer to establish will be aborted
--convert-links: convert links so that they work locally, off-line, instead of pointing to a website online
--limit-rate: imit download speed
--no-clobber: don't overwrite any existing files (used in case the download is interrupted and resumed)
--random-wait: random waits between download
--restrict-file-names: change which characters found in remote URLs must be escaped during generation of local filenames
--spider: wget will behave as a Web spider, which means that it will not download the pages, just check that they are there
--tries: set number of retries to number
--user-agent: identify as agent-string to the HTTP server