

Descargas FTP y mirrors de sitios web con Wget

CONSÍGUELO TODO

Wget descarga ficheros e incluso sitios web completos desde la línea de comandos. **HEIKE JURZIK**

Existen un buen número de administradores de descarga basadas en GUI que permiten a los usuarios descargar ficheros y sitios web completos. Sin embargo, pocos son tan flexibles ni potentes como la instrucción de la línea de comandos `wget`. `Wget` descarga rápidamente cualquier cosa sin tener que teclear ni indicar demasiado. `Wget` “habla” HTTP, HTTPS y FTP; puede continuar transferencias interrumpidas e incluso dispone de una función de actualización que únicamente actualiza ficheros que han cambiado.

Por Todas Partes

La sintaxis genérica para `Wget` es la siguiente:

```
wget URL
```

```
huhu@asteroid:~/test$ ls -l
total 2264
-rw-r--r-- 1 huhu huhu 12 Sep 5 13:35 link.bmp -> screenshot.bmp
-rw-r--r-- 1 huhu huhu 2313894 Sep 5 13:35 screenshot.bmp
huhu@asteroid:~/test$ gzip link.bmp
gzip: link.bmp is not a directory or a regular file - ignored
huhu@asteroid:~/test$ gzip -f link.bmp
huhu@asteroid:~/test$ ls -l
total 2276
-rw-r--r-- 1 huhu huhu 9543 Sep 5 13:35 link.bmp.gz
-rw-r--r-- 1 huhu huhu 2313894 Sep 5 13:35 screenshot.bmp
huhu@asteroid:~/test$ gunzip link.bmp.gz
huhu@asteroid:~/test$ ls -l
total 4528
-rw-r--r-- 1 huhu huhu 2313894 Sep 5 13:35 link.bmp
-rw-r--r-- 1 huhu huhu 2313894 Sep 5 13:35 screenshot.bmp
huhu@asteroid:~/test$
```

Figura 1: La forma más simple del comando `wget` ignora las imágenes embebidas y no sigue los enlaces.

La salida de la línea de comandos que imprime en la terminal permite ver qué está haciendo (Figura 1): en nuestro ejemplo, la herramienta está estableciendo una conexión a un servidor web (puerto 80 estándar) y descargando el fichero `index.html` a un directorio local, ignorando imágenes embebidas y sin seguir los enlaces. Si no se desea ver en la consola la salida de forma tan detallada, se puede especificar la opción `-q` (por *quiet*, es decir, funcionamiento silencioso). Sin embargo, cuando se le dice a `wget` que suprima la salida de mensajes de error y de información básica, conviene utilizar un término medio con la opción `wget -nv`. Ésta hará que el programa escriba una salida más corta en la consola pero que contiene alguna información.

Para decirle que siga los enlaces locales en el servidor y refleje los datos recursivamente, habrá que añadir el parámetro `-r`. Si se hace de este modo es aconsejable especificar la profundidad de la recursión. Será preciso bajar un nivel para obtener tanto los `index.html` como todos los enlaces embebidos (tales como imágenes u otras páginas HTML):

```
wget -r -l 1
www.linux-magazine.es
```

Si se configura el nivel de profundidad a `-l 2`, `Wget` extraerá los ficheros otro nivel por debajo de un primer nivel. En otras palabras, si `index.html` contiene un enlace a `images.html`, el administrador de descargas seguirá en este caso los enlaces a esta página.

Se crea una carpeta para cada URL en el disco duro local, pero este funcionamiento puede cambiarse añadiendo otra opción. Puede especificarse `-nH` (“no Host”) para guardar todos los resultados en el directorio actual.

`Wget` puede modificar los enlaces en ficheros HTML individuales. Por ejemplo, si se establece el parámetro `-k`, manejará referencias a imágenes, hojas de estilo, páginas HTML desde el mismo servidor, etc. `Wget` referencia enlaces a ficheros que ya ha descargado por medio de una **ruta relativa**, mientras que los ficheros que no han sido almacenados en el disco local mantendrán sus URLs completas.

¡Con Calma!

Si una descarga voluminosa se interrumpe, no hay ni de que preocuparse ni es necesario comenzar desde el principio. Con la opción `-c` (de *continue*, esto es, continuar) continuará desde donde lo dejó la descarga previa. No importa si el intento de descarga original se hizo usando `wget` o un administrador de descarga gráfica, la herramienta compara los fragmentos con el original y continua a partir de allí. Mientras lo hace, presenta bastante información. Un ejemplo de la salida es

```
The file is already fully
retrieved; nothing to do.
```

para archivos que ya existen en el disco duro.

En los casos en los que se descargan repetidamente los mismos datos, es aconsejable especificar la opción `-N`, con la cual comparará el tamaño y la fecha de cada fichero con la copia local:

```
$ wget -N
ftp://ftp.debian.de/debian-cd/
3.1_r0a/i386/iso-cd/debian-
31r0a-i386-binary-1.iso
...
The sizes do not match
(local 7935840) - retrieving.
```

Si no ha cambiado nada, el administrador de descarga dirá algo parecido a:

