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 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
```

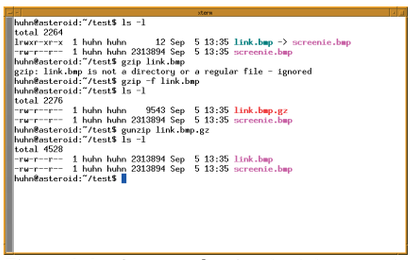

Figura 1: La forma más simple del comando wget ignora las imágenes embebidas y no sigue los enlaces.

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 empotradas 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 recursion. 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 perocuparse 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 decarga 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:

```
Server file not newer than ➥
local file "index.html" ➥
-- not retrieving.
```

Pero no hay que preocuparse si se olvida la opción: normalmente Wget no sobreescribe ficheros locales, sino que en su lugar crea un backup con un número de serie (*index.html.1*, *index.html.2*, etc.)

Especificación de Tipos de Ficheros

Si sólo quieren descargarse ficheros de un tipo específico y se intenta pasar a Wget un asterisco (*) como un **wildcard**, la herramienta presentará el siguiente mensaje de error:

```
wget ➥
www.linux-magazine.com/*.jpg
...
HTTP request sent, ➥
awaiting response...➥
404 Not Found
14:24:09 ERROR 404: Not Found.
```

En su lugar, es necesario hacer constar el tipo de fichero o una lista de tipos de ficheros a través de la opción *-A*. Por ejemplo:

```
wget -r -l 1 ➥
-A jpg,png,giff...
```

Si se observa la salida atentamente, veremos que primero descarga el *index.html*, pero luego mueve los ficheros otra vez para dejar solamente las imágenes.

También funciona en la dirección opuesta, y así, mediante el parámetro *-R*, permite a los usuarios ignorar tipos de ficheros específicos. De nuevo, el parámetro espera una lista de tipos de ficheros que no se quiere que sean transferidos al disco local. El siguiente comando

```
wget -R avi,mpg,wmv ...
```

evita esos mamotréticos archivos de video que tanto ocupan en el disco duro.

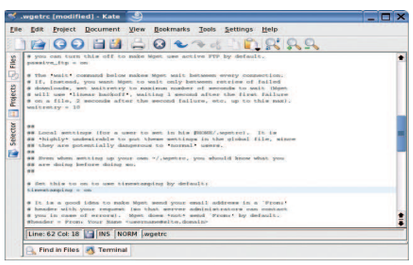

Figura 2: Este podría ser su propio "~/wgetrc". Los comentarios se indican con símbolos de almohadillas.

Frugal

Existen algunas opciones para restringir las funcionalidades de wget. Por ejemplo, si se tiene una conexión a Internet lenta y se prefiere no usar todo el ancho de banda para la descarga, puede restringirse ese ancho de banda especificando la opción *--limit-rate=*. Además, es necesario especificar el volumen en Kbytes por segundo, como en:

```
wget --limit-rate➥
=20k ...
```

El parámetro también entiende valores en Mbytes. Para 10 MBps habría que introducir:

```
wget --limit-rate➥
=10m ...
```

Si se prefiere restringir el volumen de descarga total, entonces usaremos el parámetro *-Q*. De nuevo se necesita hacer constar el volumen de datos. La opción sobreentiende valores en bytes, Kbytes o Mbytes. Por ejemplo, el comando siguiente

```
wget -Q40m
```

restringe el volumen de descarga a 40 Mbytes.

Completamente Automático

Si se tiene alguna dificultad para recordar los parámetros para wget, o si se piensa que haciéndolo de este modo es una pérdida de tiempo, puede usarse un fichero de configuración oculto en el directorio de inicio en el que definiremos nuestras propias preferencias. Para crear un fichero de configuración, hay que copiar la plantilla global */etc/wgetrc* al directorio de inicio

```
cp /etc/wgetrc  ~/.wgetrc
```

para, a continucación, arrancar un editor para modificar el fichero. El fichero tiene entradas para todas las opciones de la línea de comandos, aunque por defecto están deshabilitadas. Para configurar el parámetro *-N* por defecto, simplemente es necesario quitar el símbolo de la almohadilla (#) al principio de la siguiente línea

```
#timestamping = off
```

y reemplazar *off* con *on*. La Figura 2 permite ver un fichero de configuración de muestra para wget.

Descarga Segura

Wget también puede coger datos desde servidores donde se necesita autenticarse a través de una clave de usuario y contraseña. Para usar esta opción, hay que pasar las credenciales al programa cuando se arranca:

```
wget --http-user=➥
nombreusuario➥
--http-passw=contraseña
```

Si se hace esto, habría que cuidarse de otros usuarios fisgones: si otro usuario corre el comando *ps* para presentar los procesos activos en la máquina, el comando wget será listado junto con el nombre de usuario y la contraseña en texto claro. Como solución, podría usarse un fichero de configuración *~/.wgetrc* en el directorio de inicio. Se introduce lo siguiente, por ejemplo:

```
http_user = nombreusuario
http_passwd = contraseña
```

Y se guarda el fichero privado escribiendo

```
chmod 600 ~/.wgetrc
```

Combinación Perfecta

wget no espera la entrada del usuario, sino que sigue con el trabajo en segundo plano. Esta es una gran ventaja si se necesita correr wget en una máquina remota vía sesión SSH. Para hacerlo, primero se establece una sesión SSH y luego se arranca el programa Screen introduciendo

```
screen
```

Tras escribir el comando wget y las opciones para que comience la descarga puede presionarse [Ctrl-A], [D] para abandonar Screen. Luego puede salirse, aunque todos los procesos que han sido arrancados continuarán corriendo. La próxima vez que entremos en la máquina remota escribiremos simplemente

```
screen -r
```

para reestablecer la sesión Screen. Ahora puede comprobarse si wget realizó su tarea como se esperaba y recomenzó la descarga si fue necesario.