SelfLinux-0.10.0

wget

Autor: Johnny Graber (*linux@jgraber.ch*)
Formatierung: Torsten Hemm (*T.Hemm@gmx.de*)

GNU wget ist ein praktisches Tool, um Dateien aus dem Web zu holen. Über die zahlreichen Optionen kann man genau das erreichen, was man will; sogar abgebrochene Downloads können wieder aufgenommen werden.

Inhaltsverzeichnis

1 Der erste Einsatz von wget

`wget` wird mit jeder halbwegs aktuellen Distribution mitgeliefert. Sollte es tatsächlich nicht installiert sein, findet man es auf http://www.gnu.org/software/wget/

Der Aufruf von `wget` lautet

```
user@linux ~/ # wget [Optionen] URL
```

Will man sich das Basisrelease von Selflinux besorgen, lautet der Aufruf

```
user@linux ~/ # wget http://www.selflinux.de/basisrelease.tar.gz
```

Sollte der Download ausirgend einem Grund abbrechen, kann er mit der Option `-c wieder` aufgenommen werden:

```
user@linux ~/ # wget -c http://www.selflinux.de/basisrelease.tar.gz

--18:35:45--  http://www.selflinux.de/basisrelease.tar.gz
           => `basisrelease.tar.gz'
Auflösen des Hostnamen »www.selflinux.de«.... fertig.
Verbindungsaufbau zu www.selflinux.de[134.100.212.78]:80... verbunden.
HTTP Anforderung gesendet, warte auf Antwort... 206 Partial Content
Länge: 696,644 (noch 122,068) [application/x-gzip]
100%[==============================>] 696,644   63.75K/s  ETA 00:00
18:35:47 (63.75 KB/s) - »basisrelease.tar.gz« gespeichert [696644/696644]
```

`wget` zeigt einem alle wichtigen Optionen auf einen Blick an. Der Statusbalken zeigt, wie weit man schon vorangeschritten ist, danach folgt die Angabe der aktuellen Geschwindigkeit und hinter ETA steht die verbleibende Zeit.

Die Option `-c` ist gerade bei grossen Dateien wie ISO-Images sehr angenehm. Bricht der Download über einen Webbrowser bei 600 von 650 MB ab, ist die Datei verloren. Mit `wget` genügt das `-c` und schon wird dort weitergemacht, wo der Unterbruch statt fand.

2 Spiegeln von Webseiten

Mit `wget` können nicht nur einzelne Dateien gespeichert, sondern auch ganze Seiten gespiegelt werden. Die dafür zu verwendende Option ist `-r`. Damit wird bei der angegebene Seite rekursiv den Links gefolgt. Standardmässig folgt `-r` den Links 5 Ebenen entlang. Dabei wird nicht unterschieden, ob die Seite vom gewünschten Server stammt, oder nicht.

Die Sache mit den Ebenen ist am Anfang recht mühsam zu verstehen. Man muss sich dies wie eine gewöhnliche Sitzung mit einem Browser vorstellen. Jeder Link, den man anklickt, öffnet eine neue Ebene. Ein `-r` bedeutet also, das man von der Startseite 5 tiefer gelegene Seiten aufrufen kann.

Weist ein Link auf der 2. Seite zu  www.linux.de, wird auch dort wiederum den Links gefolgt und die Dateien auf dem eigenen Rechner abgelegt. Je nach Seiten kann dies sehr schnell mühsam werden.

Die Option `-l num` steht für "level" und passt die Tiefe von `-r` an. `num` muss durch eine beliebige positive Zahl ersetzt werden.

```
user@linux ~/ # wget -r -l 2 www.selflinux.de
```

Speichert alle Dateien, die über eine andere Datei verlinkt sind, im Verzeichnis **www.selflinux.de.** Die gefundene Verzeichnisstruktur wird dabei übernommen. Allerdings wird nur 2 Ebenen tief gesucht, was bei grossen Kapiteln dazu führt, das nicht alle Dateien heruntergeladen werden.

`wget` ist gut um sich schnell einige Seiten zu holen. Für ein effektives Spiegeln eines Servers sollte man sich ein anderes Tool suchen.

3 Das Verhalten von wget anpassen

3.1 Unterdrücken und Erzwingen von Ordnern

Bei dem Aufruf von `wget -r` wird immer ein Ordner mit dem Namen der Webseite erstellt. Will man dies verhindern, lautet der Aufruf `wget -r -nd`. Aber Vorsicht mit gleich lautenden Dateinamen: Sollte ein Name schon vorhanden sein, überschreibt `wget` den Inhalt ohne zu fragen.

Will man das Anlegen der Ordner aus irgendeinem Grund erzwingen, lautet die Option `-x` oder in der langen Version `--force-directories`. Die Verzeichnisstruktur wird nun komplett übernommen.

Neben diesen beiden bietet `wget` noch eine dritte Möglichkeit. Hierbei wird die Verzeichnisstruktur übernommen, doch wird der Ordner mit dem Domainnamen weg gelassen. Dies erreicht man mit `-nH (--no-host-directories)`.

3.2 Nur bestimmte Dateitypen herunterladen

Wildcards können bei `wget` nicht verwendet werden. Es gibt aber dennoch eine Möglichkeit, nur spezielle Dateitypen zu bekommen. Dafür muss man eine Liste mit `-A (--accept)` erstellen.

```
user@linux ~/ # wget -r -A jpg,png http://www.selflinux.de
```

Bei diesem Aufruf werden rekursiv die Dokumente nach ***.jpg** und ***.png** durchsucht und abgespeichert. Da HTTP keinen List-Befehl kennt, muss `wget` zuerst alle HTML-Dateien herunterladen, um an die Links zu kommen. Sobald die Bilder gefunden sind, werden die HTML-Dateien gelöscht.

Der Umkehrbefehl von `-A` ist `-R (--reject)`. Sollen alle Dateien, ausser ***.jpg** und ***.png** geholt werden, lautet der Aufruf

```
user@linux ~/ # wget -r -R jpg,png http://www.selflinux.de
```

3.3 Grössenbegrenzung des Downloads

Auf die Grösse des Downloads kann aber nicht nur über `-A` und `-R` Einfluss genommen werden, sondern auch mittels `-Q (--quota)`. Die Grössenangabe erfolgt in Bytes und legt den Wert für den ganzen Download fest. Da die Angabe grosser Werte in Bytes mühsam ist, kann man auch andere Einheiten verwenden. Für Megabytes wird an die Zahl ein m angehängt, für Kilobytes dient ein k.

```
user@linux ~/ # wget -r -nH -Q5m http://www.selflinux.de
```

Damit werden maximal 5 Megabyte Daten von  www.selflinux.de geholt und im aktuellen Verzeichnis abgelegt. Sind weniger als 5 MB Daten vorhanden, kann `wget` ja nicht das ganze Quota ausnutzen.

3.4 Dateien vor Download auf Datum prüfen

Holt man sich öfters Daten vom gleichen Server, möchte man ja nur die neuesten Dateien herunterladen. Mit `-N (--timestamping)` veranlasst man `wget`, vor dem Download das Datum der Datei auf dem Server mit dem

der lokalen Kopie zu vergleichen. Nur wenn die lokale Datei veraltet ist, beginnt wget mit dem Download.

```
user@linux ~/ # wget -N http://www.selflinux.de
```

3.5 Verwenden eines Proxy-Servers

Will man einen Proxy-Server verwenden, genügt die Option `-Y on/off`. Dabei wird auf die Umgebungsvariable `$http_proxy` ausgelesen. Diese muss natürlich gesetzt werden:

```
user@linux ~/ # export http_proxy="http://meinproxy.provider.de:3128"
```

4 Die Dateien /etc/wgetrc und .wgetrc

Eine grosse Anzahl der Startoptionen können in diese Konfigurationsdateien eingetragen werden. Die Datei **/etc/wgetrc** gilt für alle User, die **~/.wgetrc** nur für den jeweiligen Benutzer.

Hier ein kleines Beispiel des Aufbaus einer solchen Datei:

.wgetrc

```
###
### Sample Wget initialization file .wgetrc
###

# You can set retrieve quota for beginners by specifying a value
# optionally followed by 'K' (kilobytes) or 'M' (megabytes).  The
# default quota is unlimited.
#quota = inf

# The "wait" command below makes Wget wait between every connection.
# If, instead, you want Wget to wait only between retries of failed
# downloads, set waitretry to maximum number of seconds to wait (Wget
# will use "linear backoff", waiting 1 second after the first failure
# on a file, 2 seconds after the second failure, etc. up to this max).
waitretry = 10

# You can lower (or raise) the default number of retries when
# downloading a file (default is 20).
#tries = 20
```

5 Übersicht der wichtigsten Optionen

-V	--version
-h	--help
-c	--continue
-N	--timestamping
-r	--recursive
-o	--output-file=datei
-i	--input-file=datei
-q	--quiet
-v	--verbose
-Y on/off	--proxy=on/off
-Q2m	--quota=2m
-nd	--no-directories
-nH	--no-host-directories

Für die ganze Liste siehe `man wget`

6 Grafische Frontends

Bei all seinen Optionen ist `wget` ein idealer Kandidat für ein Frontend. Mit `gtm` und `kwebget` gibt es zwei bekanntere Programme. Da die Bedienung mit grundlegenden Kenntnissen von `wget` problemlos möglich ist, wird hier auf eine detaillierte Einführung verzichtet. Für programmspezifische Infos schaut man sich am besten die den Tools beigelegte Hilfe an.

* gtm: http://gtm.sourceforge.net/
* kwebget: http://www.kpage.de/de/