Tölts le teljes weboldalt wget-tel

LinuxOS Magazine – 2016. január

Írta: Paul Arnote (parnote)

Szerettél volna valaha is online talált információkat menteni, megőrizni, de túl nehéznek és időrablónak találtad az egyes oldalak letöltését? Vagy esetleg lélekölőnek találtad az egyes oldalak kinyomtatását, akár papírra, akár PDF fájlba? Vagy kellett már mentened egy, vagy több weboldaladat?

Semmi gáz, a wget megmenthet téged. A wget-tel teljes weblapokat, vagy csak részét letöltheted. Be kell vallanom, hogy a cikk „ötletét” a Linux Journal ezen 2008-as cikke adta.

VIGYÁZAT! NE HASZNÁLD az eljárást nagy fájlokat tartalmazó, vagy nagyon nagy weboldalakon! Nagy valószínűséggel ki fogsz fogyni a tárolókapacitásból. Emellett, sokáig tart azon fájlok letöltése – és még tovább a nagy fájloké. Még az olyan oldalak is, mint a PCLinuxOS Magazine-ém több mint 2 GB adatot és fájlt tartalmaznak. Szintén, NE foglald le ugyanazt a weboldalt újra, meg újra. Továbbá, bizonyos tartalmakat nem szabad letölteni (pl. jelszó fájlok, bankkártya információk, stb.) és „helytelen internetes viselkedés” olyan adatok, vagy tartalmak piszkálása, amikhez normálisan nem lenne hozzáférésed.

A figyelmeztetések után, próbáljunk némi alapvető ismeretet szerezni a wget-ről magáról és használatáról.

A wget parancssori eszköz (hé, ne bátortalanodj el, ha grafikus srác, vagy csaj lennél). Ha beírod a **wget --help**-et terminálba, akkor ez lesz az első, amit látni fogsz:

Usage: wget [OPTION]... [URL]...

Majd ezt követi pár ezer (szélsőségesen sok) opció.

Kétségtelen, hogy a wget RENGETEG opcióval bír, ami a wget erejét mutatja. Ugyanakkor, ezen opciók összessége egy új wget-felhasználó számára túlzás. Nem fogunk ebben a cikkben mindent érinteni, amire a wget képes. A cikk leginkább arra vállalkozik, hogy bemutassa a wget-et és megértesse annak használatát.

Akkor, most vessünk egy pillantást a wget parancsra. Ha már megnézted így egyben, akkor elemeire fogjuk szedni. A parancsot egyetlen, összefüggő sorként kell beírni.

wget -x -r -np -k -v
http://pclosmag.com/html/Issues/201511/ -P
/home/parnote-toshiba/Downloads/PCLOSMag/

Szedjük szét a parancsot. Természetesen a **wget** utasítással indítunk. A követő **-x** megmondja a wget-nek, hogy a könyvtár létrehozását erőltesse. A **-r** hatására a wget rekurzívan beolvassa a kiinduló alatti könyvtárakat is. Következik a **-np** opció jelentése „no parents,” mondja a wget-nek, hogy ne töltse le a hierarchiában fölötte álló könyvtárakat. A **-k** a következő opció, utasítva a wget-et a fájlba kerülő hivatkozások konvertálására a helyi felhasználáshoz**,** a **-v** bekapcsolja bőbeszédű

```
--2015-12-15 16:00:25--  http://pclosmag.com/html/Issues/201510/images/page08/code2.jpg
Connecting to pclosmag.com (pclosmag.com)|104.222.96.159|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 8636 (8.4K) [image/jpeg]
Saving to: '/home/parnote-toshiba/Downloads/PCLOSMag/pclosmag.com/html/Issues/201510/images/page08/code2.jpg'

pclosmag.com/html/Issues/201510 100%[=====================================================>]   8.43K  --.-KB/s   in 0.004s

2015-12-15 16:00:25 (2.34 MB/s) - '/home/parnote-toshiba/Downloads/PCLOSMag/pclosmag.com/html/Issues/201510/images/page08/cod
e2.jpg' saved [8636/8636]

--2015-12-15 16:00:25--  http://pclosmag.com/html/Issues/201510/images/page08/code3.jpg
Connecting to pclosmag.com (pclosmag.com)|104.222.96.159|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 96758 (94K) [image/jpeg]
Saving to: '/home/parnote-toshiba/Downloads/PCLOSMag/pclosmag.com/html/Issues/201510/images/page08/code3.jpg'

pclosmag.com/html/Issues/201510 100%[=====================================================>]  94.49K   538KB/s   in 0.2s
```

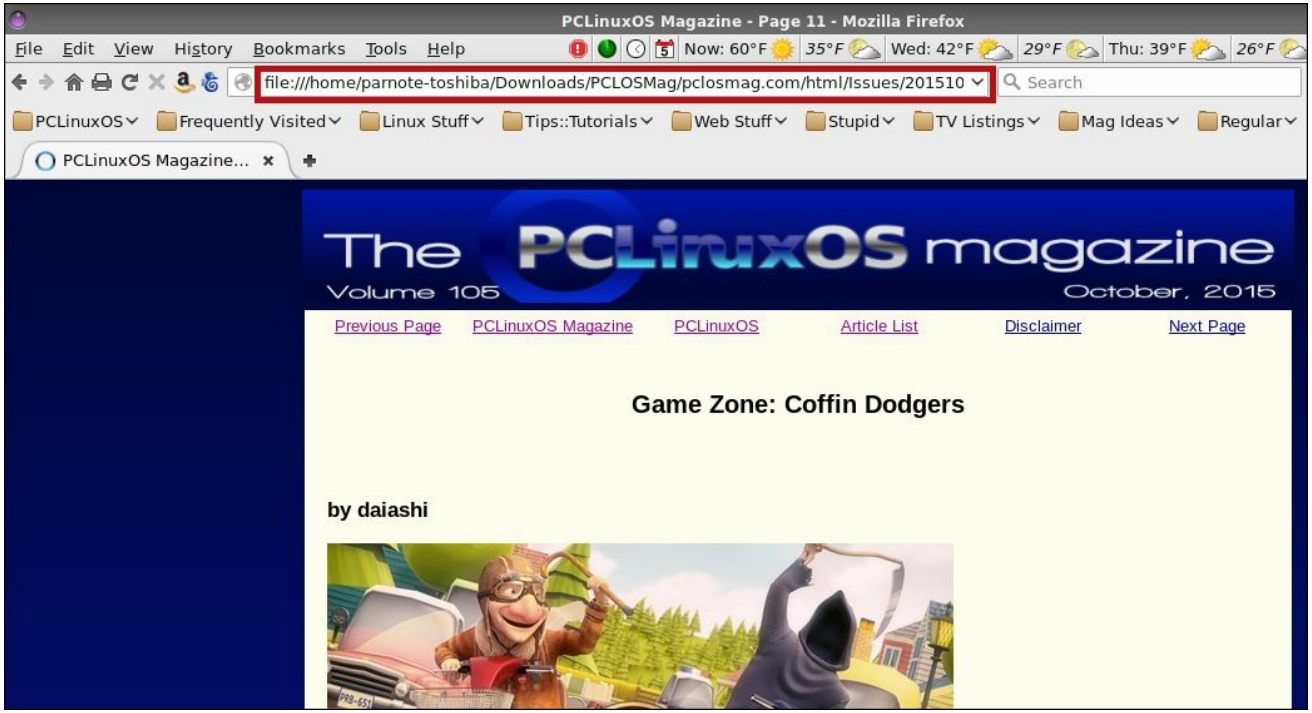
módot, ami a letöltés előrehaladásáról ad rengeteg információt.

A **http://pclosmag.com/html/Issues/201511/** opció meghatározza a weblap letöltésének kiinduló oldalát, ami megegyezik a letölteni tervezett honlapon az információt tartalmazó könyvtárral. A Magazin honlapján az egyes hónapok HTML fájljai a /html/Issues könyvtárban vannak, minden hónap egy négyjegyű év és kétjegyű hónap adattal meghatározott alkönyvtárban. Így, a 2015. decemberi kiadás HTML változatának letöltéséhez a 201511-et ki kell cserélni 201512-re, illetve a 2010. márciusi szám HTML változatához pedig a 201511-et 201003-ra cserélni.

Végül, a **-P /home/parnote-toshiba/Downloads/PCLOSMag/** opció mondja meg a wget-nek, hogy hová mentse a letöltött fájlokat. Ebben az esetben az én /home könyvtáram (/home/parnote-toshiba) Downloads könyvtárának /PCLinuxOSMag alkönyvtárába. A beírás a /home könyvtárad pontos nevétől, és attól függ, hogy hová akarod az általad letöltött fájlokat menteni.

Ez a parancs ***majdnem*** az összes, a PCLinuxOS Magazine 2015. novemberi számának HTML változatához tartozó fájlt letölti a számítógéped általad meghatározott könyvtárába. Lesznek olyan fájlok, amiket nem tölt le, mint az egyes számokban megjelenített reklámok. Ezek egy másik könyvtárban találhatók, „off limits” jelölés alatt az adatgyűjtő programok számára. Ezeket a könyvtárakat a Robots.txt fájl határozza meg a Magazin honlapján.

Van néhány wget parancsopció, ami nem használható egymással. Ilyen a **-nc** opció, ami a „no-clobber” (nincs rátöltés) helyett áll. Ez nem kompatibilis a **-k** opcióval, ami konvertálja az összes hivatkozást, hogy a kapcsolat nélküli fájlok rendben működjenek. A „no-clobber” lehetővé teszi a letöltés folytatását onnan, ahol abbahagytad, nem felülírva a már a letöltés teljes végrehajtása előtti kilépést megelőzően letöltötteket. Egy másik a **-O** opció a



Figyeld meg az URL címsort. A Firefox a helyi fájlt mutatja … és az összes hivatkozás is él!

letöltési fájl nevének megadásához (igen, a wget képes egyetlen fájlt is letölteni). Ismét csak, nem kompatibilis a **-k** opcióval. Biztos vagyok benne, hogy még vannak ezen felül inkompatibilitások, amiket még nem fedeztem fel.

Természetesen, ha a wget összes opcióját nézzük, lehetnek még további olyan parancsok, amik érdekelhetnek. Első a **-D pclosmag.com** megmondja a wget-nek, hogy ne kövesse a pclosmag,com domain-ről kilépő hivatkozásokat. Másik a **-p**, a --p-requisites helyett, megmondja a wget-nek, hogy szedje le a HTML oldal helyi megjelenítéséhez szükséges összes képet, css stíluslapot stb.-t. Harmadik a -o opció, lehetővé téve a wget számára egy log fájl meghatározását, amibe kiírja az információkat, ahelyett, hogy a terminál képernyőjén jelenítené meg. Negyedik a **--user [username]** és **--password [password]** (lecserélve a [username]-et és [password]-öt adott felhasználóra és jelszóra) átadja az adott mezőket, amennyiben felhasználót és jelszót igénylő oldalra mész. Végül a **-m** opció az egész weblapot letükrözi a helyi számítógépre, az adott kiindulási ponttól kezdve. Mindent megkapsz, kivéve azokat a könyvtárakat, amik a Robots.txt fájlban off-limit-ként vannak megjelölve.

A wget kétségkívül az online szköztár egyik nagyon hasznos eszköze. Remélem, hogy a bemutató adott némi betekintést és megmutatta a wget hatalmas erejét. Javaslom, hogy ismerd meg a wget számos további elérhető opcióját is.