



[Новости](#) • [Файловые архивы](#)
[Поиск](#) • [Активные темы](#) • [Топ лист](#)
[Правила](#) • [Кто в on-line?](#)

[Вход](#) • [Забыли пароль?](#) • [Первый раз на этом сайте?](#) • [Регистрация](#)

Компьютерный форум Ru.Board » Компьютеры » Программы » Wget

Модерирует : [gyra](#), [Maz](#)

[Версия для печати](#) • [Подписаться](#) • [Добавить в закладки](#)

Страницы: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#) [16](#) [17](#) [18](#) [19](#) [20](#) [21](#) [22](#) [23](#) [24](#) [25](#) [26](#) [27](#) [28](#) [29](#) [30](#) [31](#) [32](#) [33](#) [34](#) [35](#) [36](#) [37](#) [38](#) [39](#) [40](#) [41](#) [42](#) [43](#) [44](#) [45](#)
[46](#) [47](#) [48](#) [49](#) [50](#) [51](#) [52](#) [53](#) [54](#) [55](#) [56](#) [57](#) [58](#) [59](#) [60](#) [61](#) [62](#) [63](#) [64](#) [65](#) [66](#) [67](#) [68](#) [69](#) [70](#) [71](#) [72](#) [73](#) [74](#) [75](#) [76](#) [77](#) [78](#) [79](#) [80](#) [81](#) [82](#) [83](#) [84](#) [85](#) [86](#) [87](#) [88](#) [89](#) [90](#)
[91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [100](#) [101](#) [102](#) [103](#) [104](#) [105](#) [106](#) [107](#) [108](#) [109](#) [110](#) [111](#) [112](#) [113](#) [114](#) [115](#) [116](#) [117](#) [118](#) [119](#) [120](#) [121](#) [122](#) [123](#) [124](#) [125](#)
[126](#) [127](#) [128](#) [129](#) [130](#) [131](#) [132](#) [133](#) [134](#) [135](#) [136](#) [137](#) [138](#) [139](#) [140](#) [141](#) [142](#) [143](#) [144](#) [145](#) [146](#) [147](#) [148](#) [149](#) [150](#) [151](#) [152](#) [153](#) [154](#) [155](#) [156](#) [157](#)
[158](#) [159](#) [160](#) [161](#) [162](#) [163](#) [164](#) [165](#) [166](#) [167](#) [168](#) [169](#) [170](#) [171](#) [172](#) [173](#) [174](#) [175](#) [176](#) [177](#) [178](#)

НОВАЯ ТЕМА

СОЗДАТЬ ОПРОС

ОТВЕТИТЬ

6aDiNa



Advanced Member

[Редактировать](#) | [Профиль](#) | [Сообщение](#) | [Цитировать](#) | [Сообщить модератору](#)

[\[UNIX Shell \]](#) | [\[Рекурсивная загрузка веб-сайтов \]](#) — родительские ветки.

Wget

GNU Wget

GNU Wget – это [свободная](#) неинтерактивная утилита для скачивания файлов по HTTP, HTTPS, FTP и FTPS (и только), обладающая самым базовым функционалом загрузки одиночных файлов и рекурсивной загрузки сайтов (HTTP) и директорий (FTP).

| [Офсайт](#) | [Википедия](#) | [Фрешмит](#) | [Опен-хаб \(бывш. Охлох\)](#) | [Мануал](#) | [Ман](#) | [Ман \(русс., устар.\)](#) | [--help \(русс.\)](#) [?] |

Где взять.

Под GNU — уже стоит. Под FreeBSD — есть в портах. Под [Mac] OS X — [собрать ванильный](#).

Под Windows есть варианты: [Cygwin](#) | [GNUWin32](#) (учитывайте зависимости) | [Wget + OpenSSL by GetGnuWin32](#) | [by TumaGonx Zakkum](#) (он же [osspack32](#) и [The Moluccas](#) [?]) | [by Jernej Simoncc](#) (32 и 64 бит отдельные пакеты) | [Wget2](#).

.

Литература.

- [Рорков](#). Параметры программы wget
- [В. А. Петров](#). Wget — насос для Интернета

FAQ

Q: Можно ли простым перебором загрузить страницы (допустим) с первой по

сотую, если их адреса:

```
http://example.org/index?page=1
```

```
http://example.org/index?page=2
```

...

```
http://example.org/index?page=100
```

A: Вэ-гет не умеет делать инкрементальный перебор, поскольку это умеет делать любая командная оболочка. На Баше это делается так:

```
$ wget -E "http://example.org/index?page="{1..100}"
```

Еще раз отметьте, {1..100} — это синтаксическая конструкция Баша, а не Вэ-гета.

Эквивалентной этой будет команда:

```
$ for i in {1..100}; do wget -E "http://example.org/index?page=$i"; done
```

Или для cmd.exe:

```
for /l %i in (1,1,100) do wget -E "http://example.org/index?page=%i"
```

Q: А как собственно сайт целиком-то загрузить?

A: `$ wget -mpek "http://example.org"`

Это, наверное, самый ходовой набор ключей, но вам, может быть, более подойдут другие. Что значит каждый — легко узнать в мане.

Q: Я хочу загрузить с сайта, требующего авторизации. Что делать?

A: Проще всего кинуть куда-нибудь файл с нужными (но лишние не помешают) куками в нетскэйповском формате, затем воспользоваться ключом `--load-cookies`.

`$ wget --load-cookies cookies.txt бла-бла` # файл cookies.txt в текущей директории

У Файрфокса куки в требуемом виде можно получить, воспользовавшись расширениями «[Export Cookies](#)» либо «[Cookie.txt](#)»; у Хрома — «[Cookie.txt export](#)»

Q: Не-ASCII символы сохраняются в именах файлов как %D0%A5%D1%83%D0%B9 (или того хуже), хотя локаль юникодная.

A: Укажите ключ `--restrict-file-names=nocontrol,unix` или `--restrict-file-names=nocontrol,windows` соответственно.

Если у вас Windows и локаль не юникодная — используйте вариант от Alex_Piggy с ключом `--local-filesystem-encoding=ENCODING`, где ENCODING — имя кодировки локали в терминах iconv. Текущую локаль cmd.exe можно проверить при помощи команды `chcp`. Для русской кириллицы в Windows обычно используется CP866.

Q: Известно, что можно приказать Вэ-гету обновить ранее загруженный файл, если на сервере новее или иного размера (ключ `--timestamping`, он же `-N`). Можно приказать учитывать только дату, но не размер (`--timestamping --ignore-length`). А можно ли учитывать только размер, но не дату?

A: При помощи одного только Wget'a — нет. Возможна обработка получаемых заголовков файла при помощи средств командной оболочки. [Пример для cmd.exe](#) [?].

Q: Можно ли приказать Вэ-гету докачать файл, но только если он не изменился.

A: Нет, нельзя. Сочетание ключей `-cN` (`--continue --timestamping`), как можно было бы предположить, нужного эффекта не даст — «докачает» даже если файл изменился — получите в итоге мусор.

Q: Можно ли при рекурсивной загрузке ограничиться только ссылками, содержащими параметр `lang=ru`, т.е. грузить:

```
http://example.org/index?lang=ru
```

```
http://example.org/page?id=1001&lang=ru
```

```
http://example.org/file?id=60&lang=ru&format=dvi
```

и не грузить:

`http://example.org/index?lang=en`
`http://example.org/about?lang=fr`
и т.д.
A: Для версий < 1.14 нет такой возможности.
Общий вид URI: <протокол>://<логин>:<пароль>@<хост>:<порт>/<путь>?<параметры>#<якорь>. Так вот ключи `-I (--include-directories)` и `-X (--exclude-directories)` относятся только к пути, но не к параметрам.
В версиях > 1.14 возможно при использовании ключей `--accept-regex` / `--reject-regex`. Пример: `--reject-regex "lang=[^r][^u]"`

Q: Можно ли средствами Вэ-гета ограничить перечень загружаемых файлов по дате модификации (новее чем, старше чем)?
A: Нет такой возможности.

Q: Можно ли при рекурсивной или множественной загрузке произвольно задать целевые пути и/или имена файлов на основе пути/имени по-умолчанию (применить транслитерацию, отбросить хвостовую часть) или хотя бы независимо (сгенерировать случайно или по счетчику)?
A: Нет.

Q: То — нельзя, это — невозможно. Почему все так плохо?
A: Потому что Вэ-гет на настоящий момент — базовая программа, предоставляющая только самый базовый функционал. Если вы уперлись в потолок ее возможностей, просто смените ее на другой инструмент. Из неинтерактивных свободных программ наиболее функциональными будут:
`aria2c` [?] — для загрузки одиночных файлов по HTTP(S), FTP, бит-торренту;
`httrack` [?] — для рекурсивной загрузки («зеркалирования») веб-сайтов;
`lftp` [?] — для работы по FTP, FTPS, SFTP, FISH, а также с листингами, отдаваемыми по HTTP(S) (пример).
`curl` [?] — для работы с одиночными файлам по HTTP(S), FTP(S) и многими другими протоколами на более низком уровне.
`wput` [?] — клон wget для аплоада файлов на удаленные FTP(S) сервера.
`axel` — клон wget для многопоточной загрузки одиночных файлов по протоколам HTTP(S) и FTP(S). Порты для Windows: [2.4](#), [2.16.1](#)

Разное .
• [GUI для Wget'a](#)

Смело правьте и дополняйте шапку, однако не забывайте отписываться об исправлениях и сохранять исходный вариант под #.

Всего записей: 1551 | Зарегистр. 17-06-2003 | Отправлено: 13:39 08-11-2003 | Исправлено: anyamer, 11:40 25-12-2023

cabron666



Advanced Member

[Редактировать](#) | [Профиль](#) | [Сообщение](#) | [Цитировать](#) | [Сообщить модератору](#)

<http://wget.sunsite.dk/>
оно?

Жизнь - это рояль, клавиша белая, клавиша черная, крышка...

Всего записей: 1342 | Зарегистр. 03-02-2002 | Отправлено: 13:52 08-11-2003

GaDiNa

[Редактировать](#) | [Профиль](#) | [Сообщение](#) | [Цитировать](#) | [Сообщить модератору](#)



Advanced Member

а где там для win ?

Добавлено
кажись нашел

<ftp://sunsite.dk/projects/wget/windows/>

но что качать-то ?

Всего записей: 1551 | Зарегистр. 17-06-2003 | Отправлено: 13:56 08-11-2003

cabron666



Advanced Member

Редактировать | Профиль | Сообщение | Цитировать | Сообщить модератору

GaDiNa

Там похоже все версии выложены, см. по свежее

Жизнь - это рояль, клавиша белая, клавиша черная, крышка...

Всего записей: 1342 | Зарегистр. 03-02-2002 | Отправлено: 14:25 08-11-2003

emx



Moderator

Редактировать | Профиль | Сообщение | Цитировать | Сообщить модератору

GaDiNa

Поиск некоммерческого ПО производится в форуме Программы. Прочти правила. Переношу топик туда.

[ТА!](#)

Всего записей: 11827 | Зарегистр. 05-06-2002 | Отправлено: 23:00 08-11-2003

Activium

Junior Member

Редактировать | Профиль | Сообщение | Цитировать | Сообщить модератору

А вот еще одно исключительно полезное сабджевое местечко

Heiko Herold's windows wget spot

[_http://xoomer.virgilio.it/hherold/](http://xoomer.virgilio.it/hherold/)

Всего записей: 58 | Зарегистр. 06-11-2003 | Отправлено: 20:47 15-11-2003

Mud

Full Member

Редактировать | Профиль | Сообщение | Цитировать | Сообщить модератору

Wget Daemon - gui для wget for windows, правда бета.....
[_http://webua.net/olin/wgetdmn.htm](http://webua.net/olin/wgetdmn.htm)

Всего записей: 440 | Зарегистр. 26-09-2001 | Отправлено: 21:02 15-11-2003

popkov

Advanced Member

Редактировать | Профиль | Сообщение | ICQ | Цитировать | Сообщить модератору

Скачать самую последнюю версию и необходимые для работы программы DLL'ки можно отсюда:

<http://xoomer.virgilio.it/hherold/>

Конкретнее:

-версия 1.9.1 со справочным файлом:

<ftp://ftp.sunsite.dk/projects/wget/windows/wget-1.9.1b.zip>

- DLL'ки (для последней версии):

<ftp://ftp.sunsite.dk/projects/wget/windows/ssllibs097c.zip>

- всё вместе: <ftp://sunsite.dk/projects/wget/windows/wget-1.9.1b-complete.zip>

Добавлено
Переводы справки к программе:
<http://vap.org.ru/wget/>
<http://www.pnpi.spb.ru/~shevel/Book/node100.html>

Полезные дополнения (программы-оболочки):
Auto WGet Daemon <http://glass.ptv.ru/software/awget.html>
оболочка позволяющую работать почти без командной строки и с IE из
контекстного меню http://yugres.cjb.net/download/dl.php?file=set_iewget.exe
А если нужна прога, в которой вводишь адресок, папку, куда качать и запускается
wget с нужными параметрами, то вам сюда: shafff.narod.ru/download/wcom.zip

Всего записей: 1835 | Зарегистр. 22-03-2003 | Отправлено: 14:07 04-12-2003 | Исправлено: popkov, 20:52 13-12-2003

popkov

[Редактировать](#) | [Профиль](#) | [Сообщение](#) | [ICQ](#) | [Цитировать](#) | [Сообщить модератору](#)

Advanced Member

Как можно скачивать форумы с помощью WGET:
<http://forum.ru-board.com/topic.cgi?forum=5&topic=19&start=26> [?]

Всего записей: 1835 | Зарегистр. 22-03-2003 | Отправлено: 22:13 04-12-2003 | Исправлено: popkov, 20:49 13-12-2003

8A1eX8

[Редактировать](#) | [Профиль](#) | [Сообщение](#) | [Цитировать](#) | [Сообщить модератору](#)



Advanced Member

popkov

Цитата:

Auto WGet Daemon

Это только для OS/2 под Windows не пашет

Всего записей: 1813 | Зарегистр. 11-12-2001 | Отправлено: 23:32 04-12-2003

popkov

[Редактировать](#) | [Профиль](#) | [Сообщение](#) | [ICQ](#) | [Цитировать](#) | [Сообщить модератору](#)

Advanced Member

Параметры программы WGET

Все параметры чувствительны к регистру. Порядок их записи не важен. Они могут следовать после URL'ов, которых можно указывать сразу несколько, разделяя пробелами. Можно сокращать последовательность однобуквенных параметров, записывая их подряд без дефисов (например, "-rk" вместо "-r -k"). Для параметров, используемых с последующими аргументами, наличие пробела перед аргументом необязательно (например, записи "-o log.txt" и "-o log.txt" эквивалентны).

Поскольку параметры могут идти после URL, есть возможность явно указать, где заканчиваются параметры, и начинается URL, с помощью '--'. Например, команда

```
wget -o log.txt -- -x
```

приведёт к попытке загрузить URL "-x", записывая выводимые сообщения о неудаче в файл "log.txt".

Глобальные свойства

1) Wget всегда следует за перенаправлениями, но их не может быть больше 20. Однако они не приводят к рекурсивной выгрузке чужих сайтов, если не указан параметр '-H' в сочетании с '-r' и соответствующие сайты разрешены для загрузки.

2) Wget поддерживает cookies. Wget принимает <все> cookies, присылаемые сервером, и отправляет их ему в заголовках дальнейших запросов. Кроме того, можно заставить wget сохранять полученные cookies в файле на диске в формате, поддерживаемом Internet Explorer и Netscape. Также возможен импорт cookies, принятых этими браузерами и полное отключение использования cookies.

3) При рекурсивной выгрузке wget отправляет заголовки 'Referer'. Этот заголовок можно также задать отдельно через параметр '--referer=URL' при загрузке отдельных файлов.

4) Поддерживаются подстановочные символы в FTP-адресах и кодирование в них данных для аутентификации. Например, команда

```
wget ftp://fly.cc.fer.hr/*
```

Загрузит все файлы из корневой директории сервера <ftp://fly.cc.fer.hr/>.

5) Данные для аутентификации на <http://> и <ftp://> - серверах могут быть включены в URL:

```
ftp://user:password@host/
```

```
http://user:password@host/
```

6) Небезопасные символы в URL'ах могут быть представлены в шестнадцатеричном виде в соответствии с кодовой таблицей ASCII. Примеры небезопасных символов: "%" (представляется как "%25"), ":" ("%3A"), "@" ("%40"). Полный список таких символов см. в RFC1738 или, если в лом искать, в [этом](#) [?] посте.

7) Wget поддерживает два режима передачи файлов через FTP: двоичный режим (binary mode, **type=i** - используется по умолчанию) и текстовый (ASCII mode, **type=a**). В двоичном режиме файлы загружаются без изменений (это режим по умолчанию). В ASCII-режиме осуществляется конвертация символов конца строки между разными операционными системами. Этот режим полезен при загрузке текстовых файлов. Вот пример его использования:

```
ftp://host/directory/file;type=a
```

8) Все параметры, принимающие разделённые запятыми перечни аргументов, поддерживают правило, что указание после них пустого перечня аргументов

очищает их текущее значение (например, заданное в файле *.wgetrc*). В следующем примере вначале очищается текущий перечень исключаемых при рекурсивной выгрузке директорий, заданный в файле *.wgetrc*, а затем в него вносятся корневые папки */cgi-bin* и */dev*:

```
-X' ' -X/cgi-bin,/dev
```

Описания важнейших параметров:

-nc - не загружать существующие файлы. Удобна для продолжения закачки сайта, прерванной посередине. При этом первым делом программа будет рекурсивно обрабатывать уже загруженные файлы, не выходя в Интернет. Поэтому, в сочетании с параметром **"-rk"** этот параметр позволяет конвертировать ссылки в HTML-файлах недокачанного сайта при отсутствии подключения к Интернету (при этом, к сожалению, HTML-файлы, сохраненные с таким расширением только благодаря использованию параметра **"-E"**, будут считаться незагруженными. А сохранённые с неправильным расширением взб-страницы не будут распознаны, как HTML-файлы, и ссылки из них экстрагироваться не будут). Если не указан ни этот параметр, ни **"-N"**, то при попытке докачать недокачанный сайт все файлы будут загружены заново с перезаписью уже существующих. Этот параметр отменяет действие параметра **"-c"** при рекурсивной выгрузке.

-nd - не создавать никаких директорий, загружать все файлы в текущую директорию (если директория не задана отдельно).

-k - после окончания закачки всех файлов конвертировать все ссылки в них для локального просмотра. При этом ссылки на те файлы, которые не были загружены, будут вести в Интернет.

-E - сохранять все HTML-файлы, имеющие неправильное расширение, с расширением *.html* (очень ценная опция при загрузке сайтов на ASP, CGI или PHP, т.к. без неё получается огромное количество файлов с неправильным расширением, которые, хотя и будут после конвертации открываться в браузере при переходе по локальной ссылке, плохо распознаются системой). Однако у этой опции есть побочное действие: те файлы, расширение которых было изменено, будут повторно загружены при докачке или обновлении существующей локальной копии сайта, даже если задан параметр **"-nc"** или **"-N"**. Это связано с тем, что wget не может знать, к какому типу файла ведёт такая ссылка: *"text/html"* или *"application/binary"*? Соответственно, пока она не отправит запрос загрузки на сервер, не узнает. Тем не менее, это можно предотвратить в случае обновления существующей локальной копии сайта, если при первоначальной выгрузке использовать опции **"-kk"**, что приведёт к тому, что оригинальные неконвертированные HTML-файлы будут резервироваться перед конвертацией в файлы с расширением *.orig*. В связи с этим, если есть вероятность, что сайт придётся докачивать без конвертации скачанных страниц, необдуманное использование этой опции оказывается невыгодным: во многих случаях те взб-страницы на сайте, которые после скачивания будут иметь неправильное расширение, или совсем не нужны, или с них в любом случае не придётся начинать обзор. А ссылки на них с других страниц сайта после конвертации будут работать всё равно... Кроме того, следует помнить, что при использовании параметра **"-nc"** без **"-E"** такие файлы не только не будут загружаться заново, но также не будут конвертироваться при докачке, и вообще при докачке не будут распознаны как HTML-файлы, т.е. они будут потеряны в плане корректного отображения. А если использовать параметр **"-E"**, они при докачке будут загружены заново, обработаны и после окончания закачки корректно конвертированы... А соль в том, что, например, на сайте www.3dnews.ru все ценные файлы имеют правильное расширение. Исключение не составляют даже файлы, создаваемые по ссылке "версия для печати"... Неправильное расширение только у счётчиков: это файла с именами вида *"image.htm@count=3"*. Они все имеют одинаковый размер (3420 байт), и в некоторых папках их оказывается до 20 штук.

Поскольку сайт этот имеет размер немаленький (больше 1 Гб), за один раз закачать его почти невероятно, даже если поставить закачку на двое суток. Это связано даже не с тем, насколько у Выс быстрый канал доступа в Интернет, а с загруженностью самого сервера. Поэтому, чтобы такого рода файлы каждый раз заново не закачивать, не следует использовать параметр `"-E"` или вносить их в список запрещённых... Однако закачать сайт www.3dnews.ru до конца всё равно не получится, т.к. эти файлы счётчиков создаются по мере рекурсивной выгрузки (я дошел до `"image.htm@count=1500"`, и прекратил закачку, т.к. стало ясно, что эти файлы никогда не кончатся). После окончательного прекращения докачки, когда кроме счётчиков уже ничего не грузится, следует снова запустить её с параметром `'-nc'` в отсутствие соединения с Интернетом, и при запрете на загрузку таких файлов. Тогда все лишние файлы будут удалены, а ссылки корректно конвертированы.

-r - рекурсивная выгрузка всех файлов в вышележащих и нижележащих папках на сервере. По умолчанию, wget не заходит на другие сайты, если не указан параметр `"-H"`, однако исключением является переадресация документов с данного сайта на другие, например на сайте www.3dnews.ru в разделе "Downloads" есть много ссылок типа `www.3dnews.ru/file.rar`. Однако при переходе по ним оказывается, что `file.rar` на самом деле находится на сайте files.3dnews.ru или на ftp.3dnews.ru. Я нашел 2 способа это обойти: запретить загрузку файлов с расширением `.rar` или ещё радикальнее: отнять у текущего пользователя право на создание папок в корневой директории, где находится папка www.3dnews.ru. Для этого проще всего воспользоваться консольной командой:

```
cacls "c:\dir1\dir2" /E /P user:R
```

где `"c:\dir1\dir2"` - путь к корневой папке (текущая директория командной строки при запуске wget), `user` - имя текущего пользователя. Выполнения этой команды отключит наследование прав доступа для данной папки и удалит вообще все параметры доступа для неё, оставив только разрешение на чтение для пользователя `user`. Результатом такого действия станет следующее: wget будет посылать на сервер запрос на файл, но при попытке записи на диск каждый раз будет обнаруживать отсутствие прав доступа. При этом запрошенный файл она будет пропускать, и продолжать закачку.

-p - загружать все файлы, необходимые для корректного отображения веб-страницы. Благодаря этому параметру можно загружать с помощью wget отдельные веб-страницы целиком, включая звуковые файлы, автоматически проигрываемые при открытии веб-страницы и фреймы. Однако, по умолчанию, wget загружает только файлы, находящиеся на текущем сервере. Чтобы загружались все необходимые файлы, этот параметр надо использовать в комбинации с параметром `"-H"`, а чтобы при этом не воссоздавалась структура всех серверов, на которых расположены затребованные файлы и не загружались ненужные файлы `robots.txt`, а сам HTML-файл был конвертирован для локального просмотра, следует использовать такую команду:

```
wget -HEkp -nc -nd -e robots=off URL
```

где `URL` - адрес веб-страницы (его рекомендуется заключать в кавычки, т.к. некоторые URL'ы, содержащие небезопасные символы типа пробела, %, @, ?, неправильно интерпретируются обработчиком командной строки, если указать их без кавычек). Можно загружать веб-страницы по списку в текстовом файле, если вместо `"URL"` указать `"-i URLs.txt"`, где `URLs.txt` - имя файла со списком адресов для загрузки (по одному на строчку, в кавычки заключать нет необходимости).

-e command - выполнить команду `command` файла `.wgetrc`. После этого параметра через пробел может идти только одна команда этого файла. Такого рода команды, естественно, имеют преимущества над записанными в файле `.wgetrc`, поскольку выполняются после выполнения всех команд в нём.

-N - загружать с сервера только более новые файлы, чем уже существующие. При этом для каждого файла на сервер будет отправляться запрос, по ответу на который будет определяться время последней модификации файла на сервере и его размер. Для некоторых файлов такая информация получена быть не может,

например для файлов, генерируемых по запросу (так всегда бывает в случае веб-страниц форумов и любых других сайтов на PHP, ASP и т.п.). В этом случае wget загружает уже существующий файл повторно. При использовании этой опции в комбинации с "-kK" будут сравниваться также длины файлов. Кроме того, при использовании этого параметра полностью отключается приписывание окончаний "#" к существующим файлам. Например, если сочетать его с параметром "-nd", разные файлы с одинаковыми именами будут последовательно загружаться и перезаписывать предыдущие, т.к., при несовпадении размеров файла на сервере и на диске, существующий файл будет перезаписан независимо от даты его модификации. Однако после конвертации HTML-файлов все ссылки будут верны: ссылка в соответствующей веб-странице, ведущая к тому файлу, который на самом деле был сохранён, будет вести к локальному файлу, а во всех остальных веб-страницах - в Интернет, несмотря на то, что другие файлы тоже загружались (но потом перезаписывались).

-np - при рекурсивной выгрузке закидывать файлы только из нижележащих директорий по отношению к указанной в URL.

-Rlist. *list* - перечень окончаний имён файлов (например, *gif* или *.jpg*) и заключённых в кавычки шаблонов имён файлов (например, "*zelazny*196[0-9]**") означает загрузку всех файлов, начинающихся с *zelazny* и содержащих числа от 1960 до 1969. Можно ещё использовать "?", как один подстановочный символ). HTML-файлы будут загружаться при рекурсивной выгрузке в любом случае, но если загрузка файлов с такими окончаниями или шаблонами имён запрещена, то после анализа они будут сразу же удаляться. Пример: -

Rrar,zip,exe,"*template=print*" (запрещает загрузку файлов с расширениями *rar, zip* и *exe*, а также файлов, содержащих в имени последовательность символов "**template=print**"). Однако последние, будучи html-файлами, всё равно будут загружаться, и сразу после загрузки удаляться, как запрещённые для сохранения. В связи с этим запрещать загрузку HTML-файлов по шаблону вообще не стоит: они всё равно будут загружены, да ещё и при докачке недокачанного сайта в очередной раз будут загружены в обязательном порядке...).

-Alist - загружать только файлы со следующими расширениями: *list*. Параметр полностью аналогичен **-Rlist**, но действует наоборот.

-llevel - при рекурсивной выгрузке сайта загружать на глубину *level*, считая от стартового файла. Если указано **-l0** или **-linf**, то будут загружены все уровни (без ограничений). По умолчанию *level*=5. Под глубиной понимается удалённость последнего загружаемого файла от первого по числу промежуточных HTML-документов, которые пришлось проанализировать, не считая первый документ, и считая последний файл независимо от того, какого он типа. Таким образом, если *level*=1 и стартовая страница *index.htm* содержит ссылки на вложенную картинку *image1.jpg*, сжатый файл *doc1.zip* и страницу *download.htm*, в которой есть ссылки на *doc2.zip*, *image2.jpg* и страницу *products.htm*, то будут загружены только файлы *image1.jpg*, *doc1.zip* и *products.htm*.

-m - создать локальную копию сайта или обновить уже существующую локальную копию сайта, загружая только обновлённые файлы, сохранять листинги FTP-директорий. Этот параметр эквивалентен следующей записи: "-Nrl0-nr".

-ologfile - записывать все выводимые сообщения в *logfile*, не выводя их на экран. *logfile* - имя файла или путь к нему (абсолютный или относительный). Например:

```
wget -mkEK http://www.gnu.org/ -o /home/me/weeklog
```

-alogfile - то же, что "-o logfile", но сообщения будут дописываться в конец этого файла вместо его перезаписи.

-iURLs.txt - загрузить все файлы с адресами, указанными в файле *URLs.txt*, где *URLs.txt* - имя файла или путь к нему. В нём должно быть по 1-му URL на строку. URL'ы могут содержать данные для аутентификации. Если в качестве файла со списком URL'ов нужно использовать веб-страницу, этот параметр следует

использовать в сочетании с опциями `--force-html` и `--base=URL`.

-Xlist - не загружать файлы из следующих папок: *list*, где *list* - рзделённый запятыми список путей к папкам относительно корневого каталога сервера, где каждый путь начинается с косой черты. Например, команда **-X/cgi-bin,/people/~somebody** приведёт к тому, что каталог верхнего уровня *"cgi-bin"* и подпапка *"~somebody"* каталога верхнего уровня *"people"* не будут загружены.

Если путь к папке содержит пробелы, его можно заключить в кавычки, а всю последовательность аргументов - в апострофы.

Этот параметр можно комбинировать с параметром **"-I"**. Например, чтобы загрузить все файлы из иерархии директорий */pub*, исключая */pub/worthless*, нужно указать:

-I/pub -X/pub/worthless

-c - продолжить загрузку файла, который остался недокачанным в предыдущей сессии wget или другой программы. Это будет выполнено, только если сервер поддерживает докачку и размер файла на сервере больше, чем сохранённого на диске. Иначе, при наличии на диске данного файла, закачка произведена не будет. Этот параметр нет необходимости указывать для того, чтобы wget при внезапном обрыве соединения продолжил закачку вместо повторной загрузки с самого начала - в такой ситуации докачка, если её поддерживает сервер, является поведением по умолчанию, а параметр **"-c"** не оказывает никакого воздействия на поведение wget. Кроме того, следует быть осторожным, сочетая параметры **"-c"** и **"-r"**, т.к. тогда каждый уже существующий файл будет рассматриваться как потенциально недокачанный, и в случае модификации документа на сервере, приведшей к увеличению его размера вместо закачки новой копии он будет "докачан" т.е. будет загружена конечная часть нового файла, и дописана в конец существующего. При рекурсивной выгрузке этот параметр вступает в

8A1eX8



Advanced Member

редактирование параметром **"-N"**, что приводит к остановке загрузки или необработке некоторых HTML-файлов (этого не происходит, только если **"-N"** Фиксированный линк на стабильную версию под виндовс оказывается отключённым ввиду отсутствия в информации о дате модификации файла на сервере). Кроме того, такая простая докачка без отката может иногда давать повреждённый файл, если производится через "хромой" прокси-сервер, Всего записей: 1813 | Зарегистр. 11-12-2001 | Отправлено: 11:00 19-03-2004 | Исправлено: 8A1eX8, 11:02 19-03-2004

8A1eX8



Advanced Member

опция уже есть во FlashGet).
Редактировать | Профиль | Сообщение | Цитировать | Сообщить модератору

-Dlist - ограничить рекурсивную выгрузку следующими доменами: *list* - разделённый запятыми перечень имён доменов. Этот параметр имеет смысл только в сочетании с **"-H"**. Может употребляться вместе с обратным по действию параметром **"--include-domains -Dfst"** проверить список ссылок на файлы на скачиваемость и если можно узнать размер каждого: который действует совершенно аналогично, но наоборот: файла? Желательно с текстовыми логами.
wget -rH -Dfoo.edu --exclude-domains sunsite.foo.edu <http://www.foo.edu/> приведёт к загрузке содержимого всех серверов в домене *foo.edu*, за исключением сайта *sunsite.foo.edu* и сайтов в поддомене *sunsite.foo.edu*.

Цитата:
Всего записей: 1835 | Зарегистр. 22-03-2003 | Отправлено: 19:42 30-12-2003 | Исправлено: popkov, 06:16 08-11-2007

`-o logfile` | `-output-file=logfile`
записывать все сообщения в файл с именем logfile. Обычно такие сообщения направляются в стандартный файл сообщений об ошибках.

`i file` | `-input-file=file`
Читать список URL из файла с именем file. В этом случае нет необходимости задавать URL в командной строке. Если URL имеются в командной строке и в вводном файле, указанном параметром `-i`, то сначала будут обработаны те URL, которые находятся в командной строке. Нет необходимости, чтобы входной файл, указанной с помощью `-i`, имел вид HTML страницы. Достаточно, чтобы URL в файле были представлены в виде простого списка. Однако, вводной файл может иметь формат HTML страницы.

Если вы указали параметр `-force-html`, вводной документ будет рассматриваться как HTML страница. В этом случае возможны проблемы с относительными линками, которые вы сможете разрешить добавив в документ строку `<base href="URL">` или определив параметр `-base=URL` в командной строке.

`--spider`

Когда программа вызвана с этим параметром, то Wget ведет себя как spider, т.е. никакие документы не загружаются на локальные диски, а только проверяется, что удаленные документы существуют. Это свойство можно использовать для проверки файла `bookmarks.html`:
`wget -spider -force-html -i bookmarks.html`
Эта особенность требует значительно больше работы от Wget, чтобы вести себя функционально идентично реальному роботу типа spider.

Всего записей: 1813 | Зарегистр. 11-12-2001 | Отправлено: 01:14 06-07-2004 | Исправлено: 8AleX8, 10:22 06-07-2004

MetroidZ

[Редактировать](#) | [Профиль](#) | [Сообщение](#) | [Цитировать](#) | [Сообщить модератору](#)



Advanced Member

8AleX8

Спасибо.

Собрал вот такую команду:

```
wget -spider -o test.txt -i file.lst
```

file.lst - просто ссылки списком, без тегов.

Но wget зачем то стал скачивать самый первый файл, несмотря на параметр `-spider`.

Остановил его. Где может быть ошибка ?

А лог пишет. Хотя в нём много лишней информации (в принципе не мешает, потом какнибудь в эксель переведу, общий размер узнать и пр.):

Length: 2,972,877 [application/octet-stream]

Кстати. Повлияет на что либо если ссылки на файл вот такого вида:

<http://www.название.us/scripts/fw/ulink.asp?nr=11&f=http://www.название.com/Files/файл.exe>
?

Если убрать первую часть ссылки - вместо файла качается html.

Хотя сейчас попробовал и простые ссылки - но опять начинает качать файлы.

Причём в этих недокачанных файлах добавились вот такие заголовки:

HTTP/1.1 200 OK

Date: Tue, 06 Jul 2004 06:01:45 GMT

Server: Apache/1.3.31 (Unix)

....




Content-Type: text/plain

Rar! ;Ps

PS

Wget 1.9

Всего записей: 1795 | Зарегистр. 12-07-2003 | Отправлено: 09:58 06-07-2004 | Исправлено: MetroidZ, 10:18 06-07-2004

| | |
|---|--|
| <div>Alex_Dragon</div> <div></div> <div>Full Member</div> | <div>Редактировать Профиль Сообщение Цитировать Сообщить модератору</div> <div>А никто не знает, как заставить wget просто список url выдрать?</div> <div>Всего записей: 422 Зарегистр. 05-01-2002 Отправлено: 10:07 06-07-2004</div> |
| <div>8AleX8</div> <div></div> <div>Advanced Member</div> | <div>Редактировать Профиль Сообщение Цитировать Сообщить модератору</div> <div>MetroidZ</div> <div>Цитата:<div>wget -spider -o test.txt -i file.lst</div></div> <div>Не <i>-spider</i> а <i>--spider</i> Запусти в командной строке <i>wget --help</i> и почитай.</div> <div>Всего записей: 1813 Зарегистр. 11-12-2001 Отправлено: 10:21 06-07-2004 Исправлено: 8AleX8, 10:21 06-07-2004</div> |
| <div>MetroidZ</div> <div></div> <div>Advanced Member</div> | <div>Редактировать Профиль Сообщение Цитировать Сообщить модератору</div> <div>Цитата:<div>Не -spider а --spider</div></div> <div>на том сайте не "--" http://hepd.pnpi.spb.ru/~shevel/Book/node103.html</div> <div>оттуда и скопировал. У всех команд "-" а даже не подумал что в знаке может быть ошибка.</div> <div>Сейчас проверяю список - wget заиклился на одном из файлов:</div> <div>Connecting to messemagnet.messefrankfurt.com[193.102.59.7]:80... failed: Bad file descriptor. Retrying. Около 20 раз затем перешёл на следующий. Можно ли уменьшить кол-во попыток вот этот параметр:</div> <div>"-t number -tries=number Установить число попыток выполнить копирование равным number"</div> <div>применимо ли к данной ситуации?</div> <div>Ещё нашёл: "-Q, --quota=NUMBER set retrieval quota to NUMBER."</div> <div>Похоже это тот самый.</div> <div>PS тему бы переименовать не мешает. Поэтому в GUI сначала и спросил. Там тема проще записана "Wget" Ту тему лучше в "GUI для Wget", а эту в просто "Wget"</div> <div>----- Всё разобрался. Логи обработал - лишнее убрал спец утилитой.</div> |

Всего записей: 1795 | Зарегистр. 12-07-2003 | Отправлено: 10:56 06-07-2004 | Исправлено: MetroidZ, 15:30 06-07-2004

Mira

Редактировать | Профиль | Сообщение | Цитировать | Сообщить модератору



Advanced Member

у меня что-то не получается получить информацию о документе (HTTP)
время создания файла , кодировка и т.д.

кто знает, подскажите пзз!

заранее благодарен.

Добавлено
все вроде нашел..
правда оно в файло пишет

Всего записей: 1627 | Зарегистр. 25-11-2003 | Отправлено: 14:45 06-07-2004 | Исправлено: Mira, 14:51 06-07-2004

Mira

Редактировать | Профиль | Сообщение | Цитировать | Сообщить модератору



Advanced Member

а как быть при скачке сайта с картинками прописанными либо в ксс либо с
использованием стилей?
типа style="background-image:url("imagesr/draft/bg_logo_bottom_2.gif")
не задумывались?

Всего записей: 1627 | Зарегистр. 25-11-2003 | Отправлено: 13:13 13-07-2004

dosya

Редактировать | Профиль | Сообщение | Цитировать | Сообщить модератору



Advanced Member

Как скачать большой файл в несколько потоков?

Всего записей: 738 | Зарегистр. 27-02-2003 | Отправлено: 22:30 18-07-2004

НОВАЯ ТЕМА

СОЗДАТЬ ОПРОС

ОТВЕТИТЬ

Страницы: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90
91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125
126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157
158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178

Компьютерный форум Ru.Board » Компьютеры » Программы » Wget

Переход по форумам



Реклама на форуме Ru.Board.

Powered by Ikonboard "v2.1.7b" © 2000 Ikonboard.com
Modified by Ru.B0ard
© Ru.B0ard 2000-2024

BitCoin: 1NGG1chHtUvrtEqjeerQCKDMUi6S6CG4iC

