Basic Tips for HTTrack

This is a short user guide to help with the most basic problems, you may encounter whn trying to copy a website using the HTTrack application.

HTTrack is highly configurable. It can be preset to download several websites in one process, to stick to specific sections of websites, to include or exclude specific file types, and to include or exclude content linked to on the main website(s) being harvested.

Help for advanced settings may be found on these pages offered by the developers:

https://www.httrack.com/html/index.html

https://www.httrack.com/html/faq.html

http://forum.httrack.com/

For most users the default settings will yield satisfactory results in respect of including pages and primary content such as pictures hosted on external websites.

However, there are three basic settings that may be very helpful to be aware of, and to try as the first steps towards solving problems with copying a website:

Download speed, robots.txt setting, and browser ID.

You should always consider changing the download speed. See how and why on page 3.

Important notice: The user interface for HTTrack looks different on Windows, and on Mac. The settings and menus contain the same functions, so if you are a Mac user the Windows examples shown in this mini user guide are still valid – they will just look, and possibly be placed, a bit different. But the functions needed and the function names are the same.

Also please notice that a manual for installing and starting HTTrack on a Mac is offered on **NetLab's HTTRack page**.

List of functions

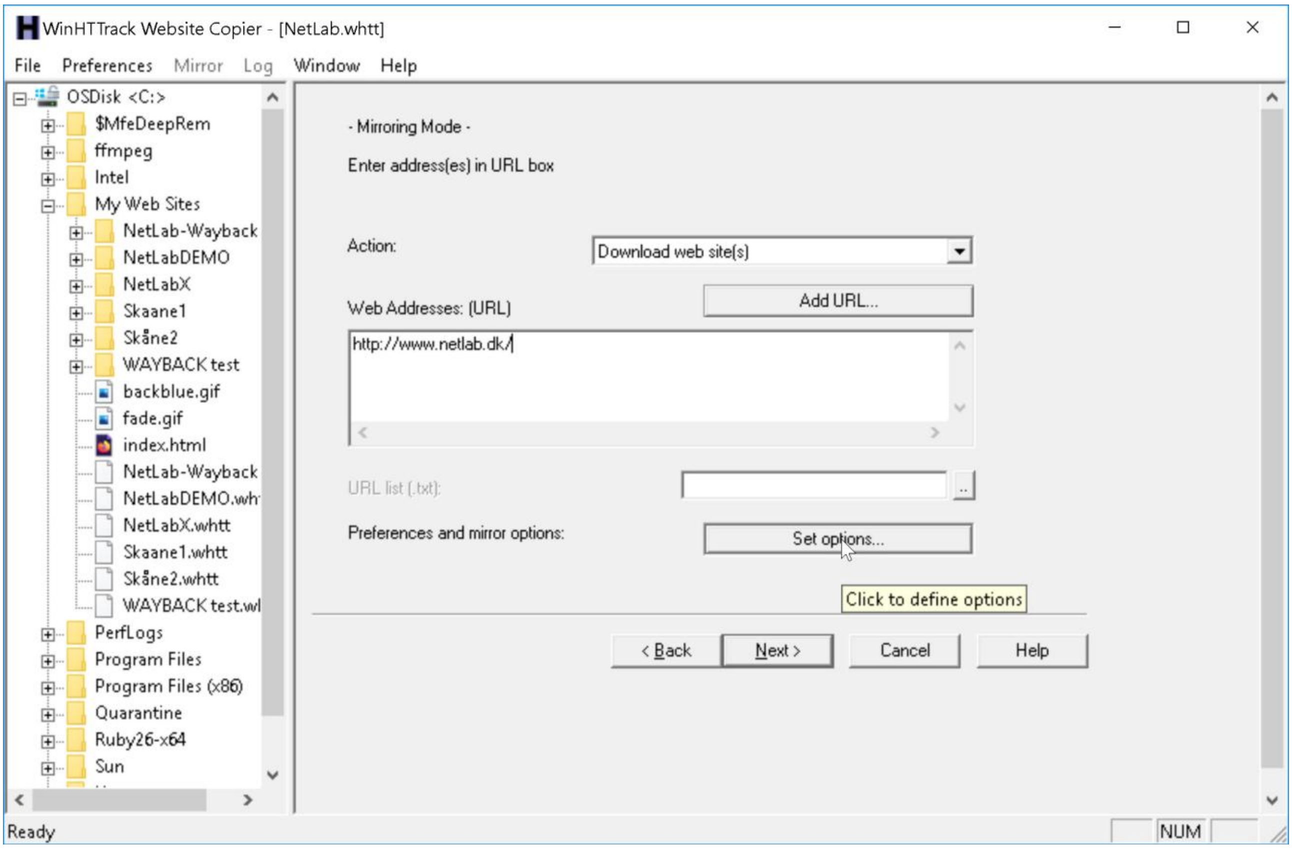
Starting a job

Before you can adjust settings for a job, you have to start one. You will have to enter a project name, you may also specify a project category (free text entry, and fully optional), and press “Next”.

This will lead you to the application window where you must specify the website(s) to be crawled.

You can now click “Set options” in order to specify changes to the default settings.

Important: “Set options” may not appear until you have entered the website URL and proceeded to the next screen on a Mac.



WinHTTrack Website Copier - [NetLab.whtt]

- Mirroring Mode -

Enter address(es) in URL box

Action: Download web site(s)

Web Addresses: (URL) http://www.netlab.dk/

URL list (.txt):

Preferences and mirror options: Set options...

When you press “Set options” a menu for advanced settings will appear, as shown on the following pages.

Change download speed

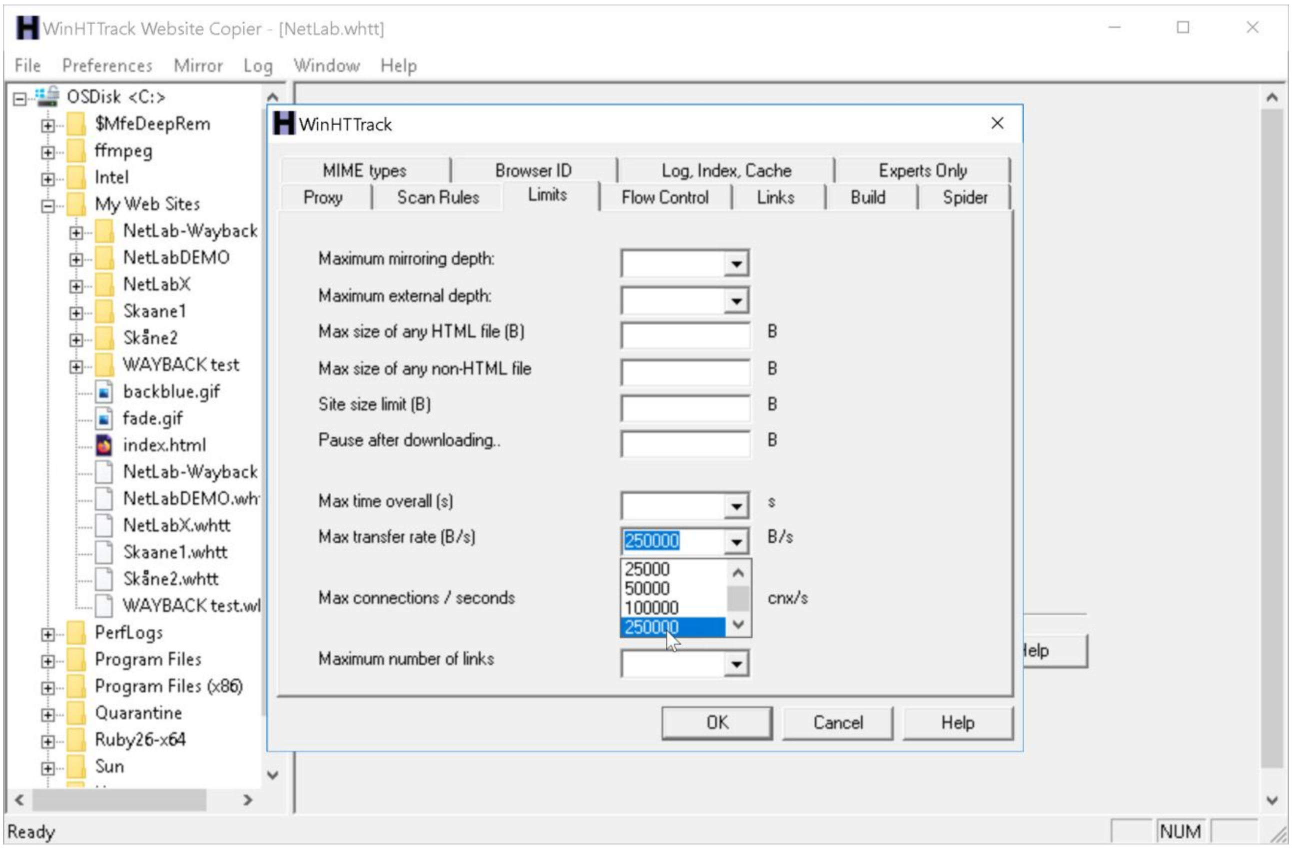
HTTrack has inherited and old setting designed for early ADSL lines with relative low capacity and speed. The default setting for how much connection capacity the program may use is very low, so as not to disturb your connection unduly.

However, modern broadband settings are usually so fast, that you will not notice any changes in download, streaming, or surfing speeds for other applications if the HTTracks download speed is set to maximum.

The only change you are likely to notice is that HTTrack will process websites significantly faster. This is in you best interest, since the full process of downloading a website may take many hours, and the process may need to run overnight – even at maximum speed setting.

In order to get the maximum setting:

In the tab "Limits", change "Max Transfer rate (B/s) from the preset 25000 to the maximum value (250000).



Maximum mirroring depth:

Maximum external depth:

Max size of any HTML file (B) B

Max size of any non-HTML file B

Site size limit (B) B

Pause after downloading.. B

Max time overall (s) s

Max transfer rate (B/s) 250000 B/s

25000
50000
100000
250000

Max connections / seconds cnx/s

Maximum number of links

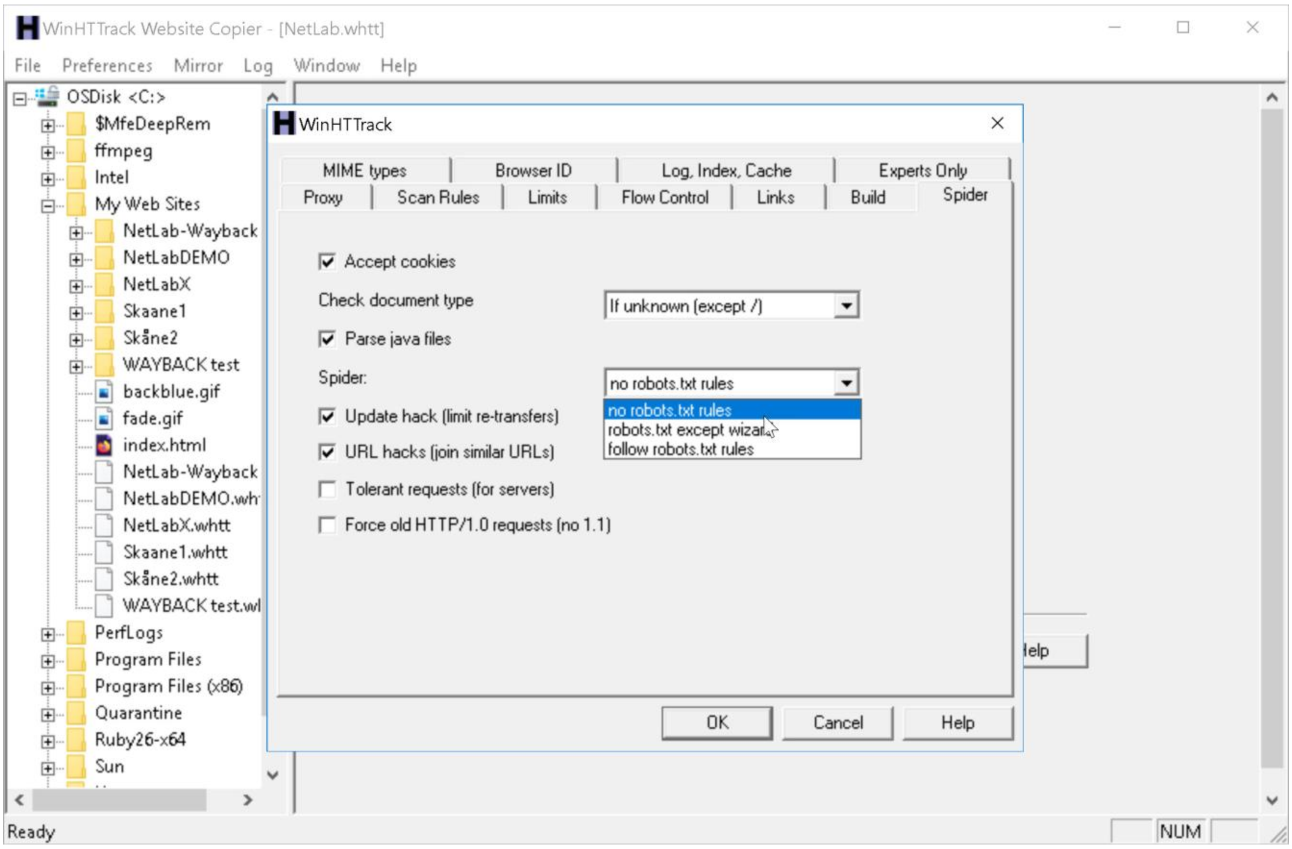
Ignore “Robots.txt”

If a website is not copied using the default settings, it is most likely because it is programmed to resist copying from archiving programs.

Resistant pages will almost certainly be protected by a robots.txt file. This is a small file telling webcrawlers that for some reason the website owner disencourages copying content from this specific website.

Solution: In the settings menu tab “Spider” change "Follow robots.txt rules" to "No robots.txt rules".

(You are not acting illegally in doing so; you are only acting against some safety precautions that the website owners for some reason preferred. When the website is public, then taking a copy may be compared to taking a photograph in public space.)



☑ Accept cookies

Check document type: If unknown (except /)

☑ Parse java files

Spider: no robots.txt rules

- no robots.txt rules
- robots.txt except wizard
- follow robots.txt rules

☑ Update hack (limit re-transfers)

☑ URL hacks (join similar URLs)

☐ Tolerant requests (for servers)

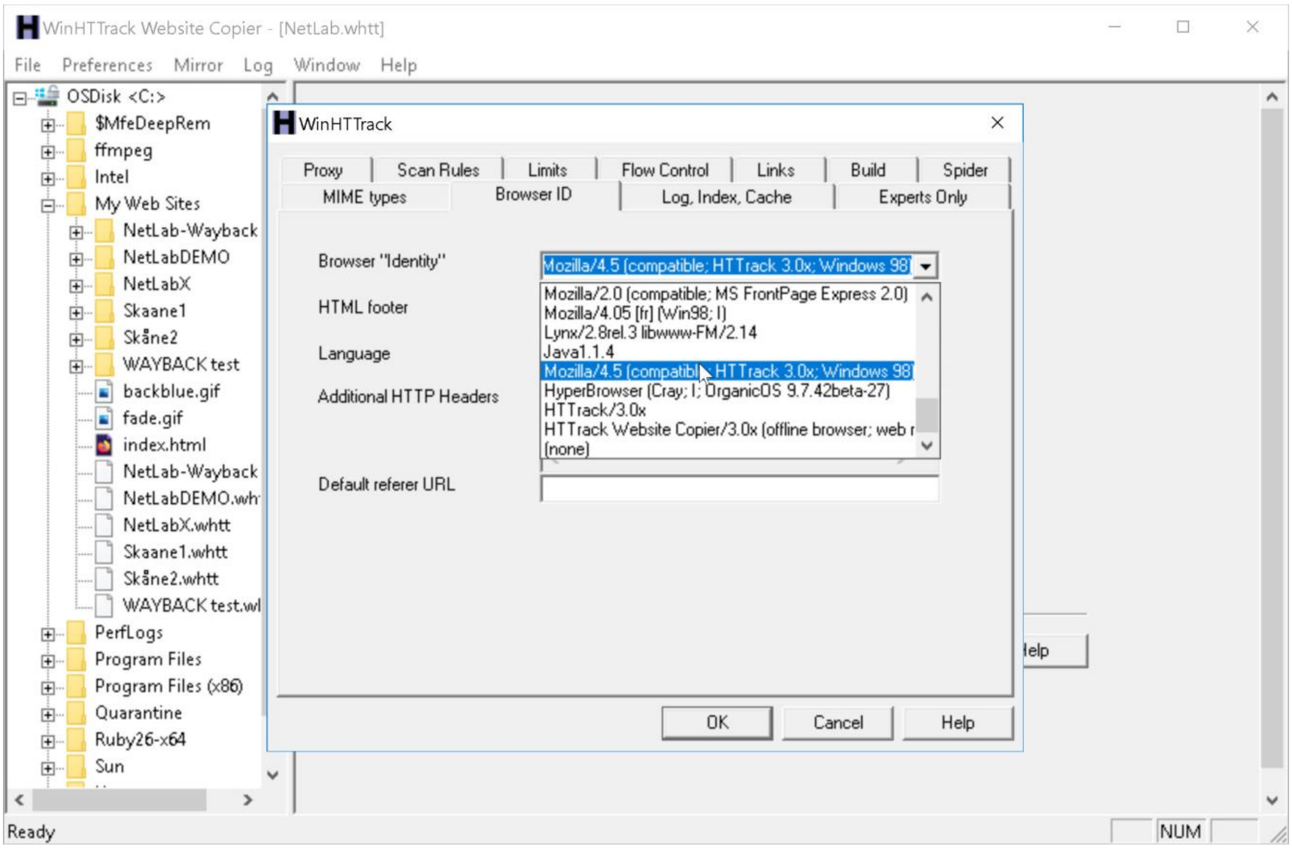
☐ Force old HTTP/1.0 requests (no 1.1)

Change Browser ID

If the website does not download now, try the tab Browser ID and have HTTrack present itself to websites as an older browser. Mozilla 4.5 compatible is usually a good option here. If a website has been programmed to recognize and reject crawlers, this will make the crawler's activity simulate a user visit.

Again, the website may have been designed with protection for various reasons, to trotect ownership, or possibly from fear of surveillance. But you have your research and archiving purposes and are only circumventing something made for security reasons that never had you or your research in mind.

Change browser ID in the tab “Browser ID”.



Browser "Identity": Mozilla/4.5 (compatible; HTTrack 3.0x; Windows 98)

It is recommend to try the latest Mozilla Firefox ID offered. If this fails, other browser IDs may still be able to solve the problem.