

## Adgang til UCloud

UCloud, der er en High Performance Computer (HPC), lokaliseret på SDU, og hvortil Aarhus Universitet har adgang.

UCloud er et sikkert computermiljø, der kan bruges til at opbevare og behandle de data, der arbejdes med i faget.

Denne manual beskriver, hvordan man bruger sit drev samt programmerne på UCloud. For yderligere instruktion kan henvises til instruktionsvideoen 'UCloud - Gephi and Orange with Ubuntu VDE' nedenfor, der også kan findes på YouTube, [UCloud - Gephi and Orange with Ubuntu VDE](#). Der omtales i videoen en folder — 'dat' — den hedder i UClouds system 'data'.



### 1. Adgang

Du får adgang til UCloud med dit WAYF. Når du logger på første gang, skal du i nederste venstre hjørne finde dit ID-nummer. Det skal videreformidles til underviseren samme med hold- og læsegruppenummer. Herefter sender underviseren en bekræftelsesemail, og du har nu adgang til din læsegruppes drev.

## 2. Datafiler

Du kan lægge dine data i mappen 'Data' på dit drev, og du kan også lave de undermapper heri, som du har behov for. Drev/mapper findes ved at trykke på 'Files' i venstremenuen.

Du uploader dine data ved at bevæge dig hen til den ønskede mappe, trykke 'Upload', og så drag/droppe.

## 3. Start af analysesoftware

Der er lavet adgang til to softwares, henholdsvis Gephi til netværksanalyse samt Orange til tekstanalyse.

Både Gephi og Orange kører på det, der hedder en virtuel disk, dvs. at den desktop, du ser programmerne i, forsvinder igen, når du lukker programmet. Det er derfor *vigtigt*, at du husker at *gemme* dine data i din data-mappe (eller undermappe), inden du lukker programmet/den virtuelle desktop, ellers er data væk.

Åbning af Gephi og Orange er stort set ens, dog med en enkelt lille forskel til slut. Her følger en trinvis beskrivelse for begge, efterfulgt af det, der er særligt for hver (brug evt. også ovenstående instruktionsvideo).

Man kan have både Gephi og Orange åben samtidig.

### 3.1. At åbne Gephi og Orange

- klik Apps i venstremenu
- find Ubuntu Xfce, marker den evt. som Favorite, så ligger den fremover øverst
- klik på den
- Job name udfyldes
- Hours, sæt fx til 3
- Machine type, vælg u1-standard-4
- Initialization: klik ned til mappen 'init' på dit drev (klik på disk-ikonet for at få alle mapper frem i projektet) og vælg det script, som svarer til det program, man vil køre (hhv. gephi-`vd.sh` og orange-`vd.sh`)

- Select additional folders to use: klik Add folder, Klik ned til 'data'-folder + evt. undermapper, hvor dine data ligger
- start programmet ved at klikke Submit, hvorefter der skrives en masse linjer
- når der står '[INFO] You can now press open Interface...' klikkes på knappen Open interface øverst til højre; hvis den laver Timeout, så tryk reload
- der åbner nu et nyt browservindue med den virtuelle desktop (hvis ikke åbner med det samme, så Reconnect, evt. flere gange)

### 3.2 Særligt for Gephi

- øverst til venstre på desktoppen ligger et ikon 'Gephi', dobbelt-klik på det, og Gephi åbner
- Dine datafiler åbnes ved i filåbneren at gå tilbage til roden (mappe/), her vælge ned i mappen 'work', og så er man nede i den mappe, man tidligere specificerede ovenfor under 'Select additional folders to use'
- filerne SKAL hentes og gemmes der, idet den virtuelle desktop som nævnt *ikke* gemmer filerne, de er væk inde i den, når den lukkes — husk det, når filerne skal gemmes, så skal man lige manøvrere ned i 'work'-mappen igen
- herefter kører Gephi som normalt

### 3.3 Særligt for Orange

- åbnes via Terminal, så den skal du lige finde ikonet for på desktoppen
- i Terminalvinduet skrives 'python -m Orange.canvas' + tryk Enter, vær omhyggelig med at skrive det korrekt
- Orange starter nu op i et nyt vindue (kan tage lidt tid)
- ved Orange er der allerede i menuen til venstre en mappe, der hedder 'work', heri kan datafilerne ses

## 4. LibreOffice

UCloud har også et gratis alternativ til Office-pakken, LibreOffice, som har en næsten identisk kopi af Excel, der hedder Calc.

Du får adgang til den i Apps:

- klik på Apps i venstremenuen
- find LibreOffice og gør den til Favorite, og klik på den
- Job name udfyldes
- Hours, sæt fx til 3
- Machine type, vælg u1-standard-4
- Select additional folders to use: klik Add folder, Klik ned til 'data'-folder + evt. undermapper, hvor data ligger
- start job programmet ved at klikke Submit, hvorefter der skrives en masse linjer
- når der ikke skrives flere linjer, klikkes på knappen Open interface øverst til højre; hvis den laver Timeout, så tryk reload
- der åbner nu et nyt vindue med programmet

## 5. Support

Hvis der er brug for support i forbindelse med selve UCloud (men ikke til Gephi og Orange), så email til frontoffice AU, [chcaa@cas.au.dk](mailto:chcaa@cas.au.dk).

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 [CDMM's HTTRack page](#).

## List of functions

Starting a job .....	6
Change download speed .....	6
Ignore "Robots.txt" .....	8

Change Browser ID ..... 9

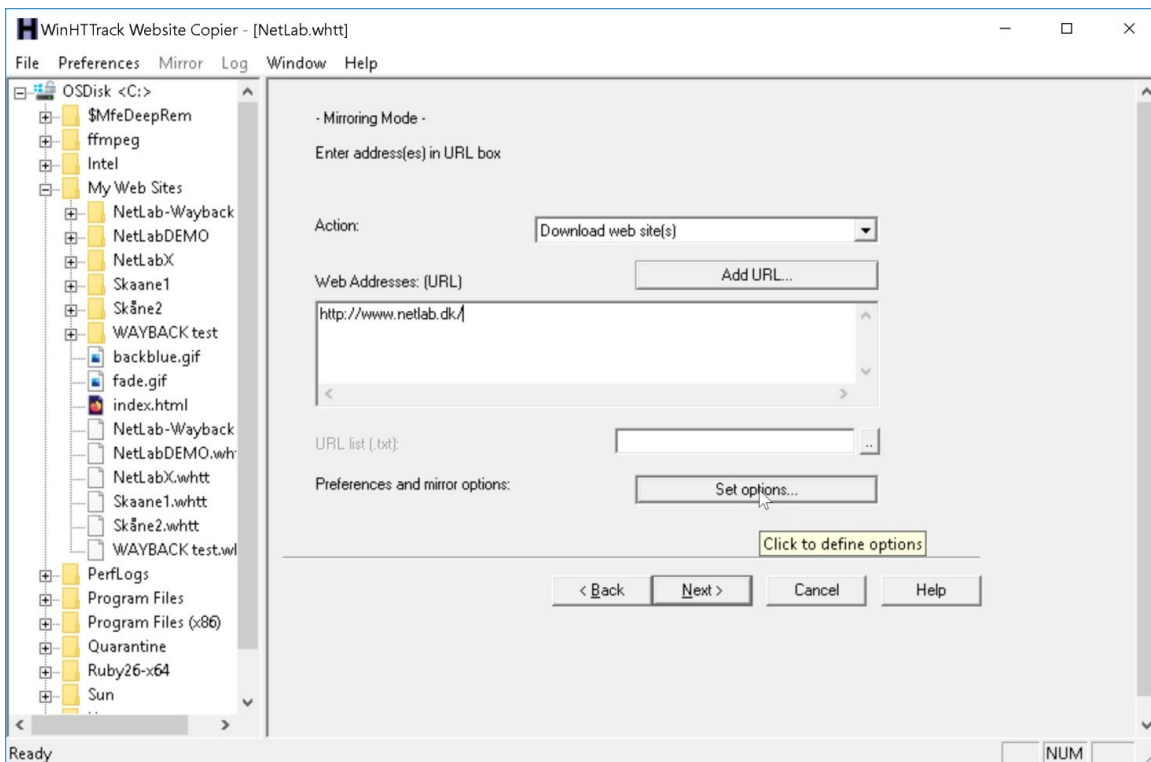
## Starting a job

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.



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

## Change download speed

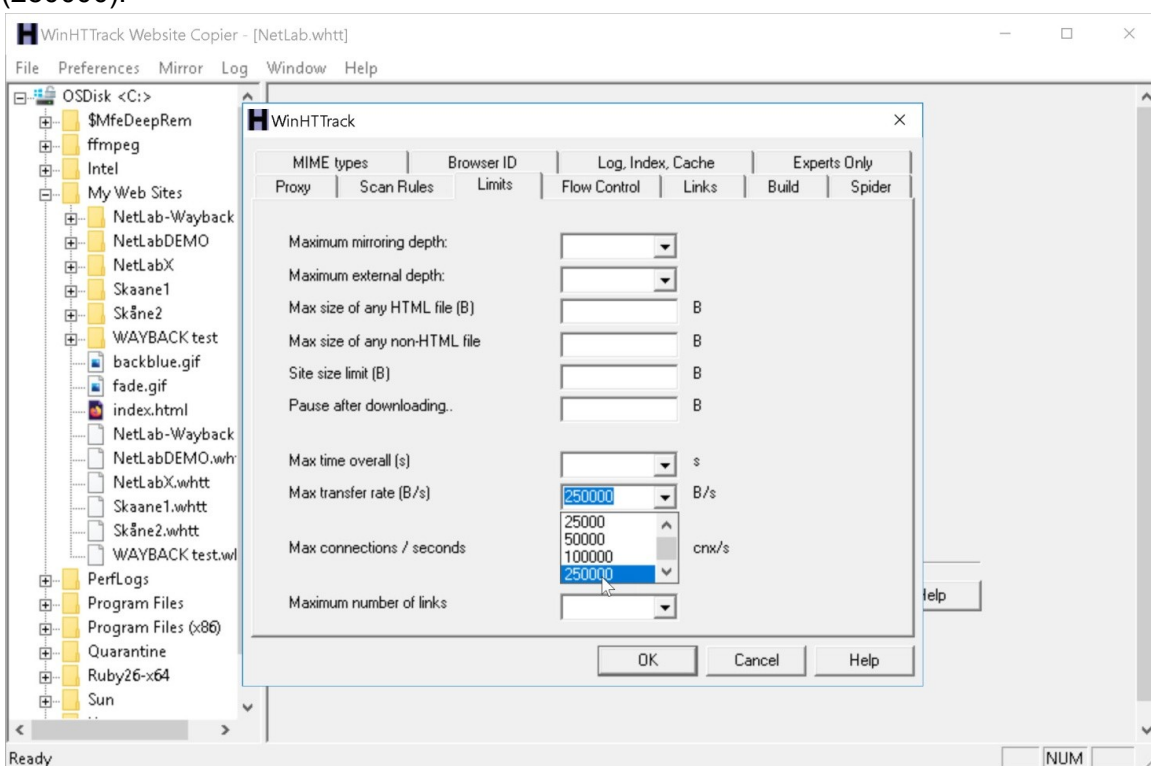
HTTrack has inherited an old setting designed for early ADSL lines with relative low capacity and speed. The default setting for how much connection capacity the program may use is very low, so as not to disturb your connection unduly.

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 your 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).



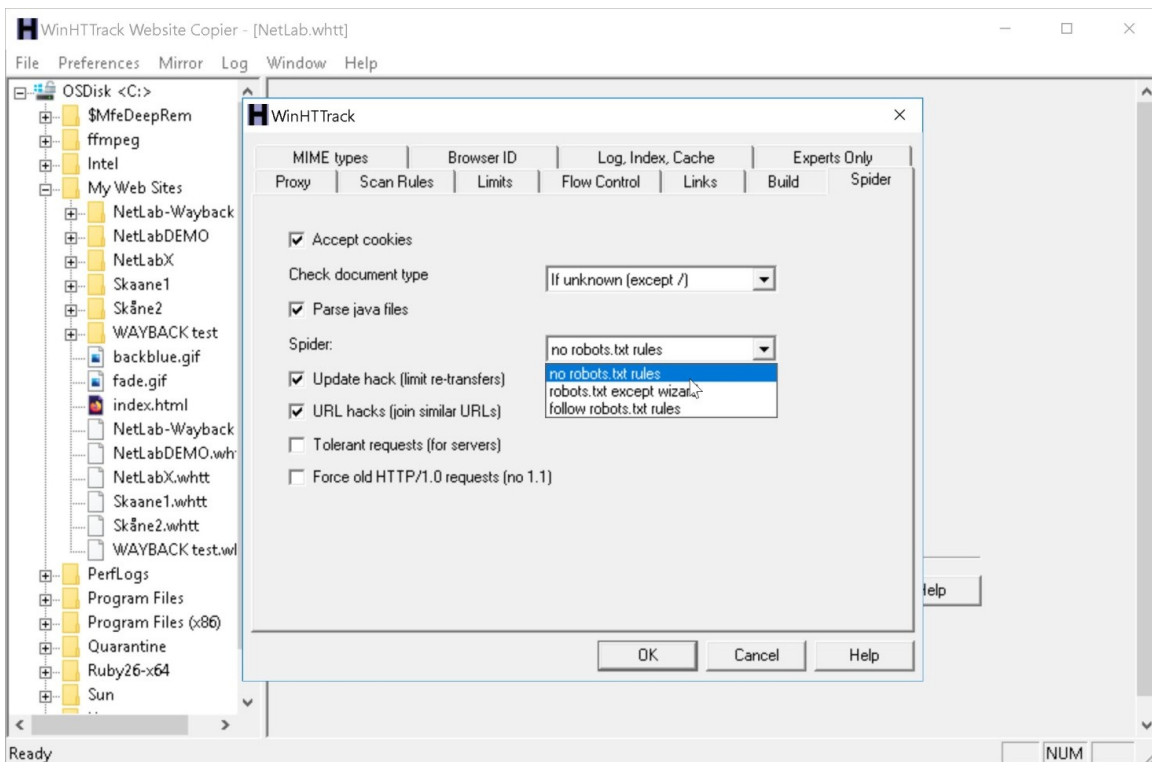
## Ignore “Robots.txt”

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 web crawlers that for some reason the website owner discourages 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.)



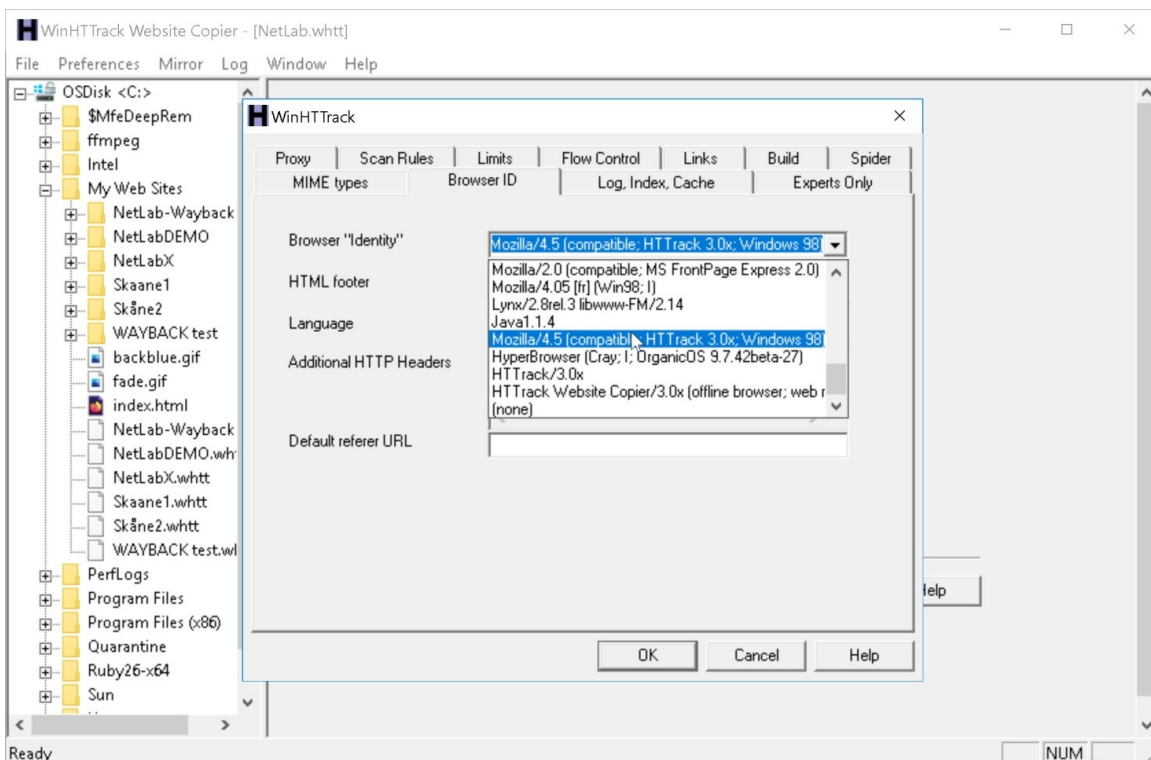


## Change Browser ID

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 protect 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".



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