InterPARES 3 Project
International Research on Permanent Authentic Records in Electronic Systems
TEAM Canada

Title: Case Study 09 – Alma Mater Society of the University of British Columbia: Policies and Procedures for Web Site Preservation

Workshop 02 Action Item 25 – Preservation Process / Strategy

Status: Final (public)

Version: 1.3

Date Submitted: September 2008

Last Revised: May 2013

Author: The InterPARES 3 Project

Writer(s): Helen Callow
School of Library, Archival and Information Studies,
The University of British Columbia

Project Unit: Research

URL: http://www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_wks02_action_25_v1-3.pdf

Document Control

Version history			
Version	Date	By	Version notes
1.0	2008-09-30	H. Callow	Discussion draft prepared following identification of action items for CS09 at TEAM Canada Plenary Workshop 02.
1.1	2008-11-07	R. Preston	Addition of Action Item text; minor content and copy edits.
1.2	2008-11-09	H. Callow	Minor content edits following feedback from R. Preston.
1.3	2013-05-27	R. Preston	Minor content and copy edits for public version.

Action 25: S. Goldfarb, with assistance from the Graduate Research Assistants assigned to case study 09, to research the best process/strategy for preserving the archival content of the AMS Web site and propose a recommendation to TEAM Canada (L. Duranti)[1]

The University of British Columbia (UBC) Alma Mater Society (AMS) is a student-run organization. Therefore, it sees mass turnover of the majority of staff on an annual basis. As a result there is no systematic method for saving records, especially those created for Web publication throughout the organization.

Some paper documentation is sent to the archives through informal agreements between the archivist and the students in charge of the records, and this tradition is passed from student to student as they come and go in and out of the organization. In this way, the archivist for the AMS receives records from the students that are valuable to the student organization as evidence of its business practices and ongoing business workflow.

The content of the Web site of the organization is another matter entirely. As more and more records are created electronically and posted to the Web, the archivist of the Society sought to preserve the Web site to ensure as much information was stored for future research.

Many changes to Web content are made in incremental stages. Currently, changes to pages evolve over a process that includes e-mails, drafts on paper and face-to-face conversations until all parties agree on a final live page. This incremental process, however, occurs on the live site with minor changes made over time until satisfaction is reached. Action Item 23 addresses this issue; and procedures have been drafted to ensure that negotiations take place on a staging site before the page is put up live for public viewing.[2]

Because there is no systematic method for saving records posted to the Web site for the long term, and because the nature of the organization sees frequent leadership turnover, the AMS has two choices in the preservation of the Web site. The first is through direct transfer of the files, through the relatively stable position of the AMS Web Editor, which are then burned onto DVD-ROM and indexed on the AMS Web Server. The second option is for a remote harvester to crawl the site and to archive the material onto the AMS Web Server, followed by a mandated disc burn.

The appraisal decision has been made to capture the entirety of the four gigabyte Web site on the first round of capture. The capture method used would determine the frequency and amount of capture. If the site is captured manually, by direct transfer, the whole of the Web site should be captured each time to limit time spent on each capture. Under this option, it is recommended that direct transfer should occur at an interval of every six months. However, if the automatic remote harvest option is used, the Web site could be captured initially as a one-time whole, and then the site could be crawled on a more frequent basis as the harvester could be set to only capture pages that have changed since the last crawl.

[1] InterPARES 3 Project, "TEAM Canada Plenary Workshop #02: Action Items and Decisions," 4. Available at http://www.interpares.org/rws/display_file.cfm?doc=ip3_canada_wks02_action_items_v1-2.pdf.

[2] See: Helen Callow (2008), "Case Study 09 – University of British Columbia Alma Mater Society: Workshop 02 Action Item 23 – Procedures for Updating Web Site Content." Available at http://www.interpares.org/rws/rws_research_studies_documents.cfm?cs=9.

Direct Transfer Option

Direct transfer works by acquiring a copy of the data, in this case the AMS Web site, directly from the original source. This requires direct access to the host Web server. Direct transfer then involves copying the selected files from the server and transferring them to the collecting institution. To ensure continued functionality, minor adjustments may need to be made to the archived site.[3] This is a viable option for the AMS and one that the Web Editor has proposed.

Remote Harvest Option

The environment for capture is the AMS server, which runs on a Windows 2003 platform. This limited the choice of a cost-free remote harvester to the HTTrack created by Xavier Roche. HTTrack is a free and easy-to-use offline browser utility. It allows a user to download a Web site from the Internet to a local directory, building recursively all directories, copying HTML, images and other files from the server to the local directory. HTTrack arranges the original site's relative link-structure. It allows users to simply open a page of the "mirrored" Web site in their browser and to browse the site from link to link, as if viewing it online.[4] This harvester has been used successfully by archivists seeking to preserve Web content in the Microsoft / Windows environment similar to the technological environment in which the AMS server operates.[5]

One advantage of this type of automatic capture is that the harvester can be installed on the AMS Archivist's computer, allowing him full control over the process to ensure capture. The primary drawback to this method is that the Content Management System (CMS) application's Expression Engine that the AMS uses has a feature to prevent unwanted access (spidering or remote harvesting).[6] It was unclear at the time that this report was written whether the CMS allows changes to this "feature" to allow crawls to be undertaken.

Storage

Whichever capturing method is used, the archived Web site needs to be preserved and stored on a relatively stable electronic digital medium. Currently, no electronic digital medium can be considered archival due to concerns regarding the relatively short and/or unproven life spans of such media and to concerns regarding technological obsolescence resulting from rapid changes in the technological environment. Therefore, whichever medium is chosen for storage will need to be periodically checked and/or refreshed to counteract data loss.[7] It is therefore

[3] For example: The hyperlinks within the archived site may need to be adjusted from absolute links to relative links; and the appropriate search engine (the one used in the original environment) must be installed in the new environment to ensure that search functionality is preserved. For a more comprehensive explanation please see: Adrian Brown, *Archiving Websites* (London: Facet Publishing, 2006).

[4] See the HTTrack Web site for more information: http://www.httrack.com/. Also see: Appendix 1 for the HTTrack User Manual.

[5] For a recent discussion of implementation, see: Christopher J. Prom and Ellen D. Swain (2007), "From the College Democrats to the Falling Illini: Identifying, Appraising, and Capturing Student Organization Websites," *American Archivist* 70(2): 344-363.

[6] The HTTrack remote harvester has been tested on numerous Web sites outside of the AMS Web site and has been determined to be a suitable candidate for use providing that changes to the CMS are made to allow for a crawl to be undertaken.

[7] See The National Archives of the UK's Digital Preservation Guidance Note: 2, "Selecting Storage Media for Digital Preservation" Authored by Adrian Brown, Head of Digital Preservation Research, August 2008. Available at: http://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf Last accessed September 29, 2008.

recommended that the archived AMS Web site be stored in several environments—for example, on the AMS server and on DVD-R—and stored in the archives to counteract these storage concerns and help assure long-term access to the stored records.

Currently, the AMS server is primarily used to redirect users to the Whitematter server where most of the AMS Web site content is stored. The AMS server is a likely candidate for use as a storage environment. It is recommended that the files also be burned to DVD-R removable media and that they also be copied to either the Web Editor's hard drive (if the direct transfer method is used) or to the AMS Archivist's hard drive (if the remote harvester option is implemented). This ensures that the archived Web site is stored in three separate places—the AMS server, a staff member's hard drive and on removable DVD-R media—thereby reducing the risk of losing the archived files to technological obsolescence, media deterioration, theft, etc. Regarding storing the Web site on hard drives, it is recommended that new hard drives be installed in the respective machines, or that external hard drives be purchased, so that the new hard drives can be dedicated to the archival process. As cost is an issue for the AMS, a quick breakdown of cost for various hard drives has been included in this report: 300 GB external hard drives can be purchased for as little as $70 (US) and internal hard drives range from $70 (US) for a 500 GB capacity to $95 (US) for a 750 GB capacity hard drive.[8] However, the hard drives will also need to be periodically checked and refreshed, and new hard drives purchased when the old drives reach full capacity.

The National Archives of the UK's Digital Preservation Guidance Note: 2, "Selecting Storage Media for Digital Preservation" mentioned earlier in this document has compiled a scorecard that looks at various storage media in terms of their longevity, capacity, viability, susceptibility, obsolescence and cost. According to their research, the DVD-R is the most effective media in terms of the AMS's needs. DVD-R have enough storage capacity to store a 4 GB Web site and are relatively affordable and easy to use. A typical DVD-R has a capacity of 4.7 GB and a cost of around $20 for a spindle of 100 units.[9] The author of the guidance note suggests using different brands or batches of the chosen media to minimize data loss due to specific manufacturers or batches having problems. The AMS should take this recommendation into consideration when purchasing media for the storage of their archived Web site, as well as the recommendation to conduct routine, periodic inspections of the files on the storage media to check for data corruption. It is also recommended that the DVD-R media be refreshed entirely every few years until testing by standards agencies has been done to discover more completely the archival capacity of the medium.

Checks

Once the Web site has been captured and transferred to the AMS environment, checks must be conducted to ensure that all the parts of the Web site captured are working as they should. Checks include, but are not limited to: manually going through and clicking on all the hyperlinks; randomly clicking on links; or employing the use of a link testing application to help automate the checking process by testing to see that all links are working.[10]

[8] See the New Egg company Web site: www.newegg.com Last accessed September 29, 2008.

[9] Ibid.

[10] See, for example: Link Checker Pro: http://www.link-checker-pro.com/; Site Audit: http://www.blossom.com/site_audit.html; Cyber Spyder Link Test: http://www.cyberspyder.com/cslnkts1.html; Link Sleuth: http://home.snafu.de/tilman/xenulink.html.

The AMS has one or two options for the preservation of their Web site in terms of both capture and storage. Capture can be achieved by directly copying the files that make up the Web site or by an automated remote harvesting tool. In either case the Web site can be navigated just as if in the online environment. It is recommended that storage options be mirrored by having the same files stored on the AMS server, a dedicated hard drive, and by copying the files to DVD-R to guard against failure of any one of the storage environments. Storage media should be periodically checked and refreshed to avoid loss of data through corruption of the various media involved.

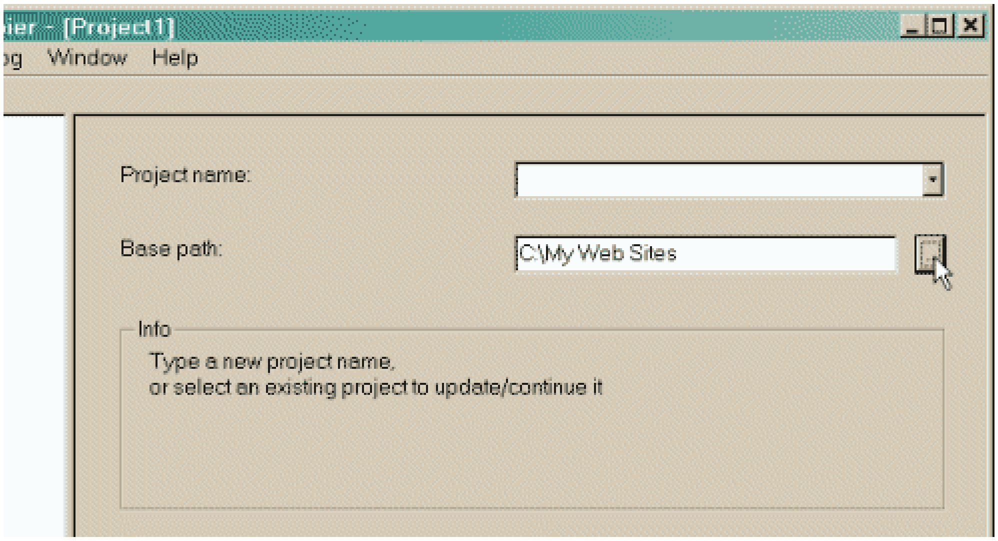
Appendix 1: HTTrack User Manual[11]

Step 1 : Choose a project name and destination folder

1. Change the destination folder if necessary

   It is more convenient to organize all mirrors in one directory, for example **My Web Sites**

   If you already have made mirrors using HTTrack, be sure that you have selected the correct folder.



Project name:

Base path: C:\My Web Sites

Info

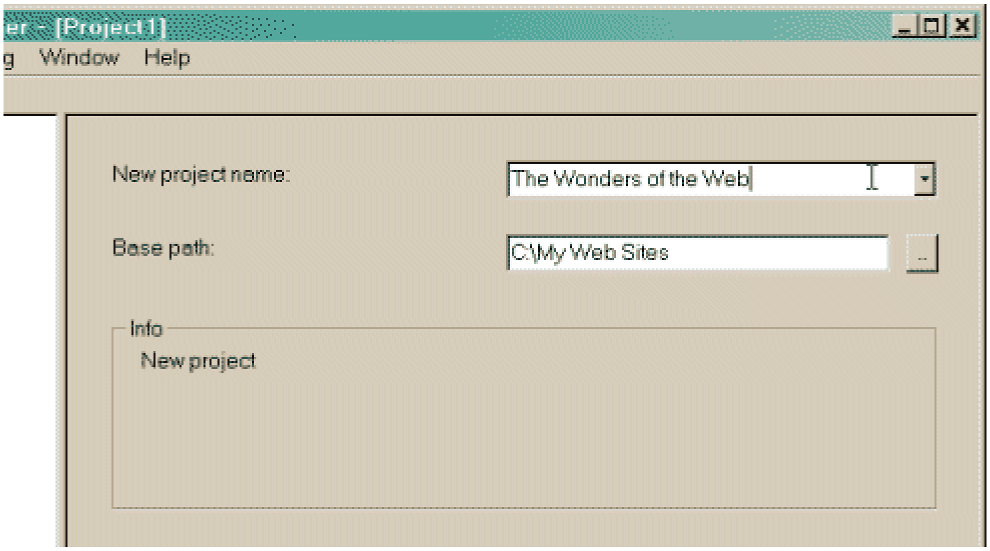
Type a new project name,
or select an existing project to update/continue it

2. Select the project name:

   - Select a new project name

     This name is, for example, the theme of the mirrored sites, for example **My Friend's Site**

[11] Copied from: http://www.httrack.com/html/step.html.



New project name: The Wonders of the Web

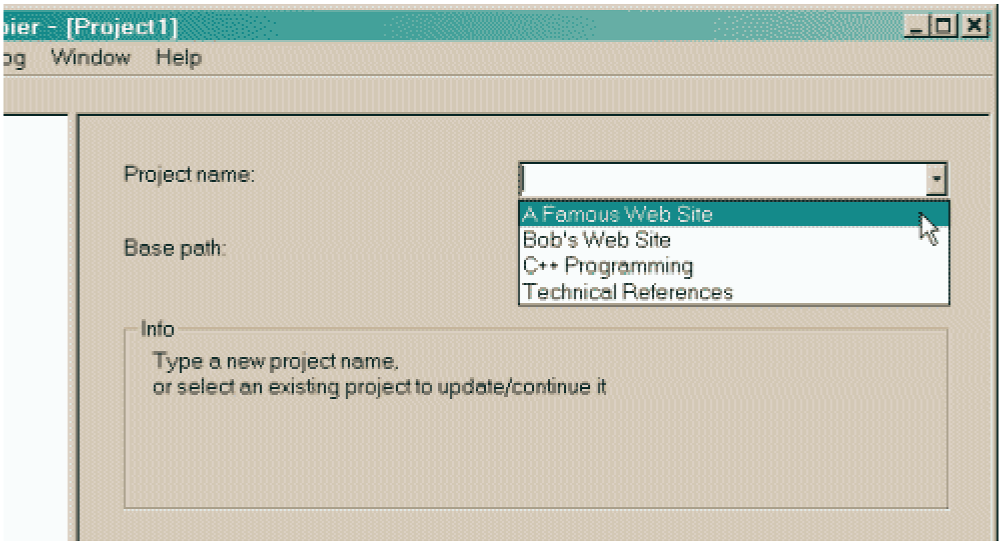
Base path: C:\My Web Sites

Info
New project

OR

- Select an existing project for update/retry

Directly select the existing project name in the popup list



Project name:

A Famous Web Site
Bob's Web Site
C++ Programming
Technical References

Base path:

Info
Type a new project name,
or select an existing project to update/continue it

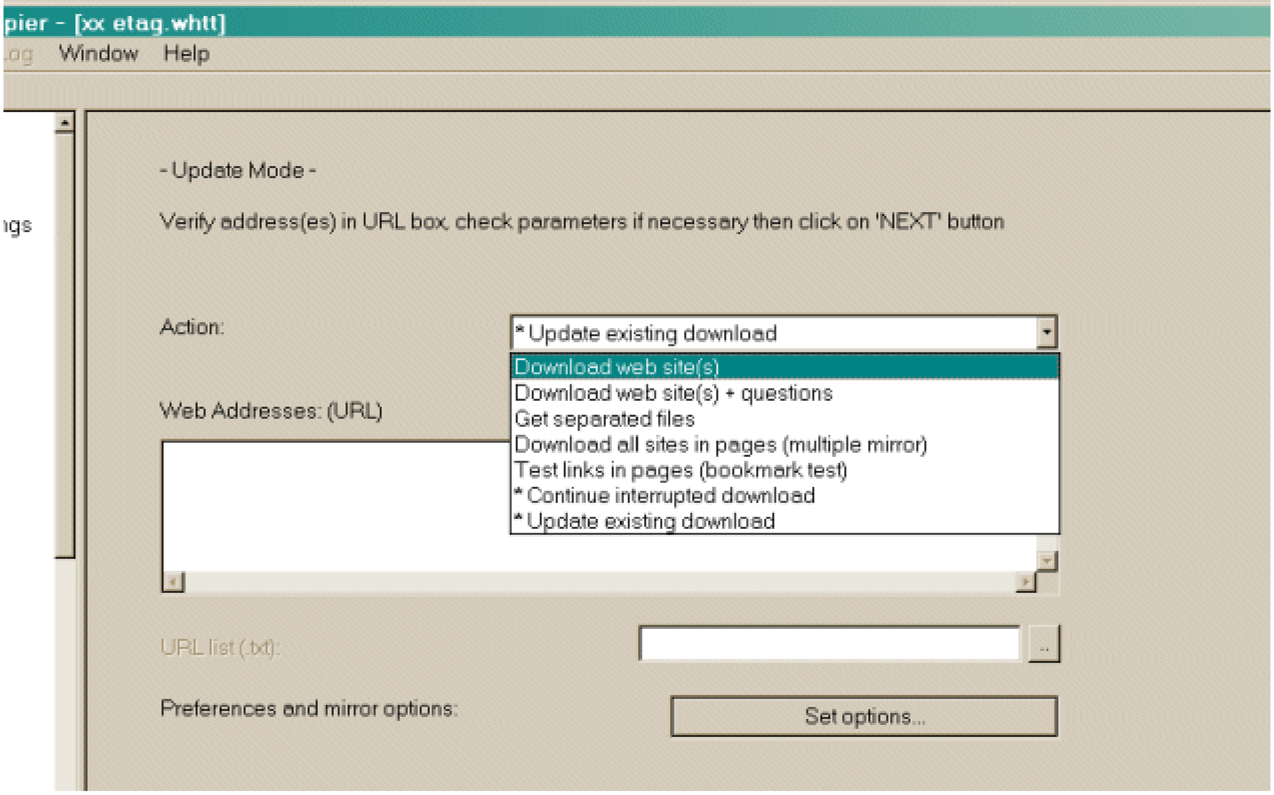
3. Click on the **NEXT** button

4. Go to the next step...

Step 2 : Fill the addresses

1. Select an action

The default action is **Download web sites**



- Update Mode -

Verify address(es) in URL box, check parameters if necessary then click on 'NEXT' button

Action: * Update existing download

pier - [xx etag.whtt]
og Window Help
ngs
Download web site(s)
Download web site(s) + questions
Get separated files
Download all sites in pages (multiple mirror)
Test links in pages (bookmark test)
* Continue interrupted download
* Update existing download
Web Addresses: (URL)
URL list (.txt):
Preferences and mirror options:
Set options...

- ○ Download web site(s)

  Will transfer the desired sites with default options

- ○ Download web site(s) + questions

  Will transfer the desired sites with default options, and ask questions if any links are considered as potentially downloadable

- ○ Get individual files

  Will only get the desired files you specify (for example, ZIP files), but will not spider through HTML files

- ○ Download all sites in pages (multiple mirror)

  Will download all sites that appears in the site(s) selected. If you drag&drop your boormark [sic] file, this option lets you mirror all your favorite sites

- Test links in pages (bookmark test)

  Will test all links indicated. Useful to check a bookmark file

- * Continue interrupted download

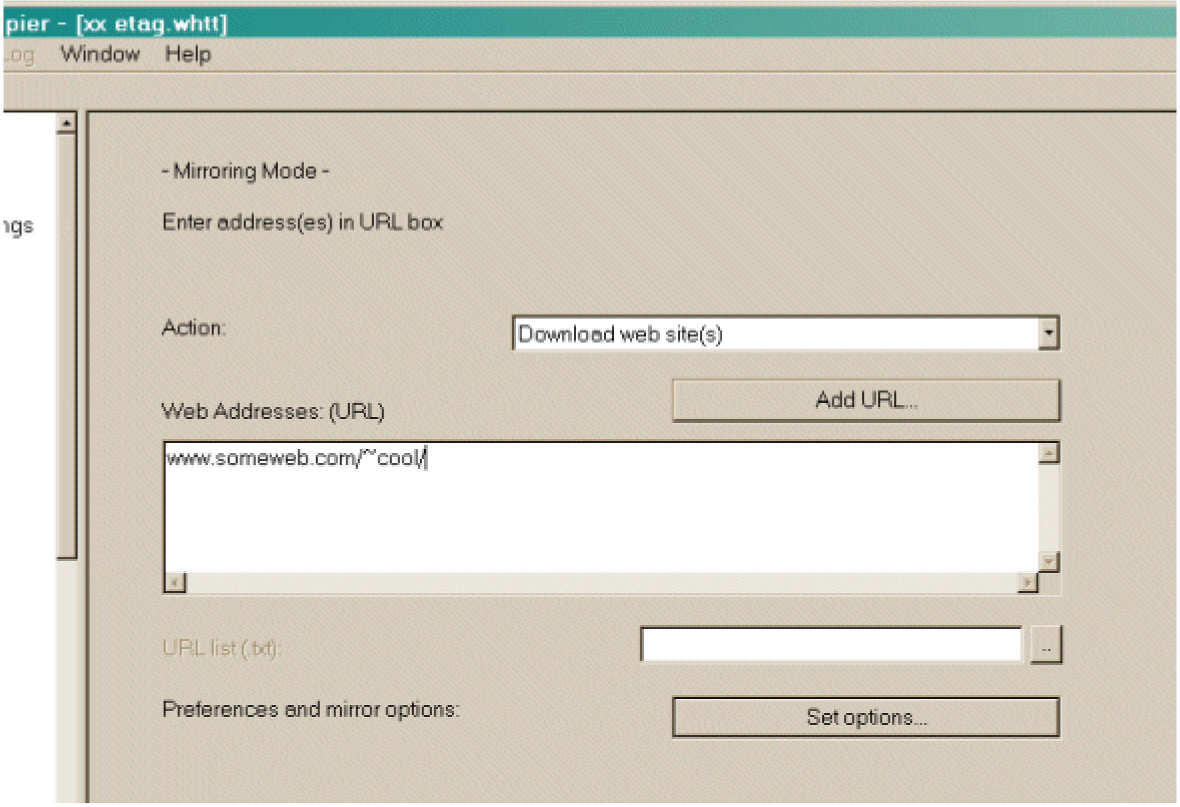
  Use this option if a download has been interrupted (user interruption, crash..)

- * Update existing download

  Use this option to update an existing project. The engine will recheck the complete structure, checking each downloaded file for any updates on the Web site

2. Enter the site's addresses

   You can click on the **Add a URL** button to add each address, or just type them in the box



- Mirroring Mode -

Enter address(es) in URL box

Action: Download web site(s)

Web Addresses: (URL)

www.someweb.com/~cool/

URL list (.txt):

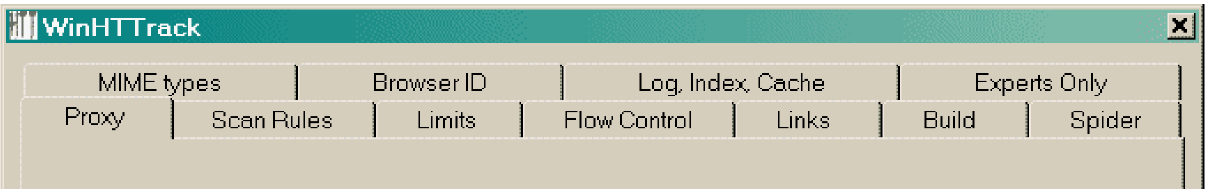
Preferences and mirror options:

3. You may define options by clicking on the **Set options** button

Option panel

- Click on one of the option tab below to have more information

Each option tab is described, including remarks and examples



You can define filters or download parameters in the option panel

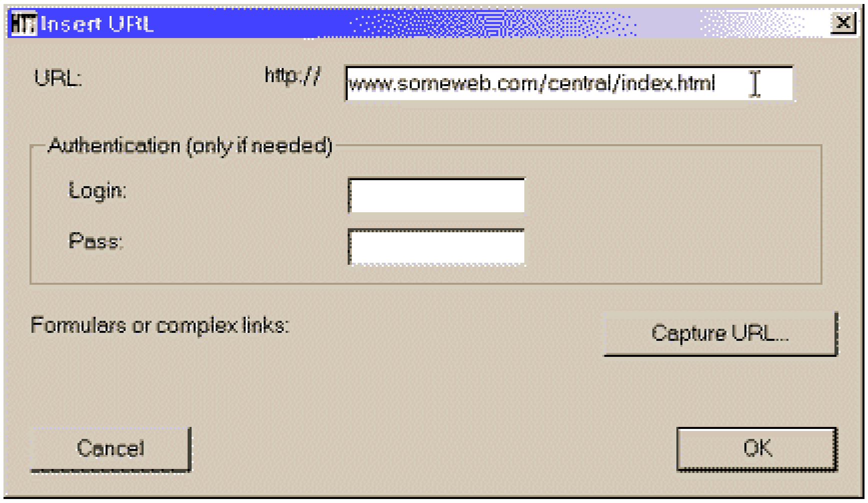
4. You may also add a URL by clicking on the **Add a URL** button

This option lets you define additional parameters (login/password) for the URL, or capture a complex URL from your browser

Add a URL

1. Enter a typical web address

Just type in your address in the field

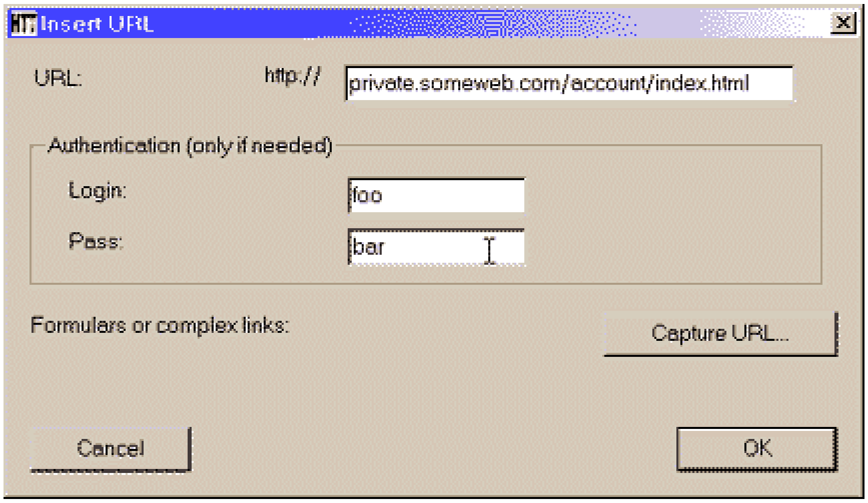


URL: http:// www.someweb.com/central/index.html

OR

2. Enter a web address with authentication

Useful when you need basic authentication to watch the Web page



URL: http:// private.someweb.com/account/index.html

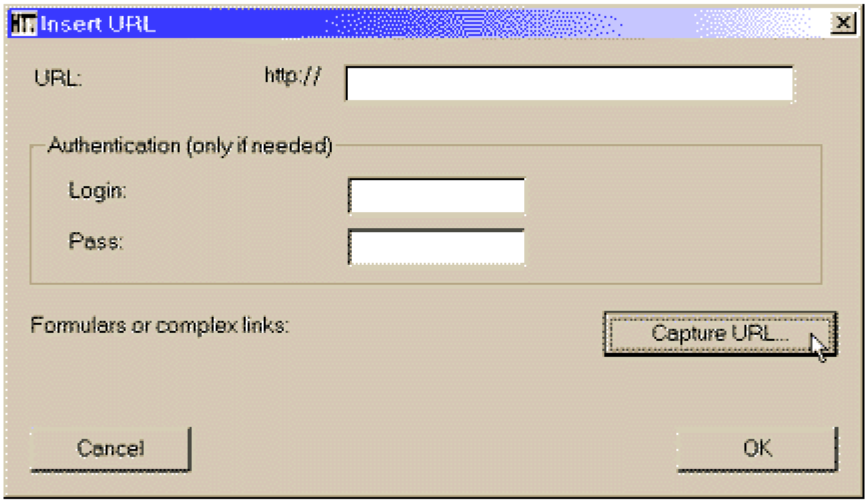
Login: foo

Pass: bar

OR

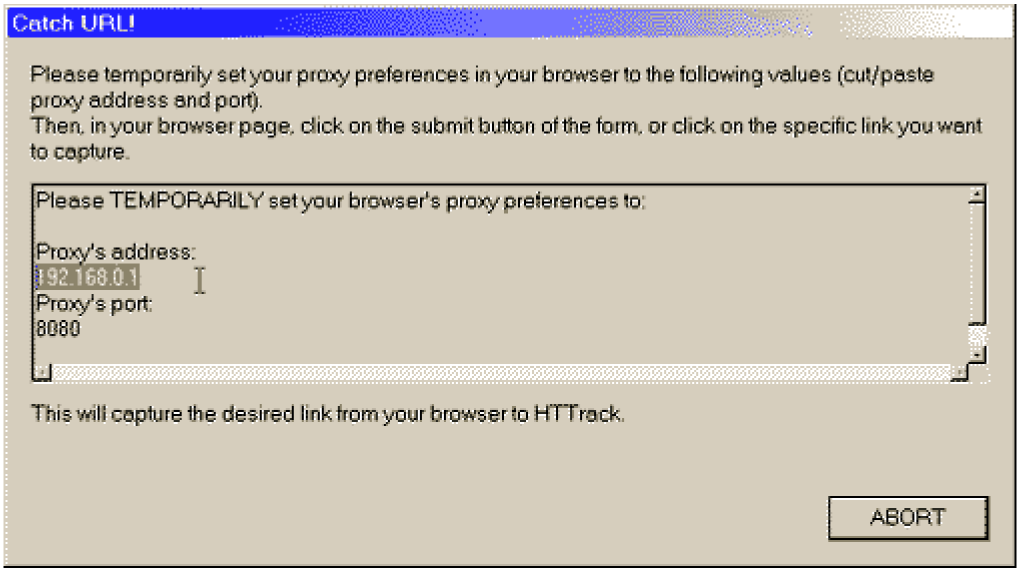
3. Capture a link from your web browser to HTTrack

Use this tool only for form-based pages (pages delivered after submitting a form) that need some analysis



Set, as explained, your Web browser proxy preferences to the values indicated : set the proxy's address, and the proxy's port, then click on the button or link as you usually do in your Web browser.

The temporary proxy, installed by HTTrack, will then capture the link and display a confirmation page.

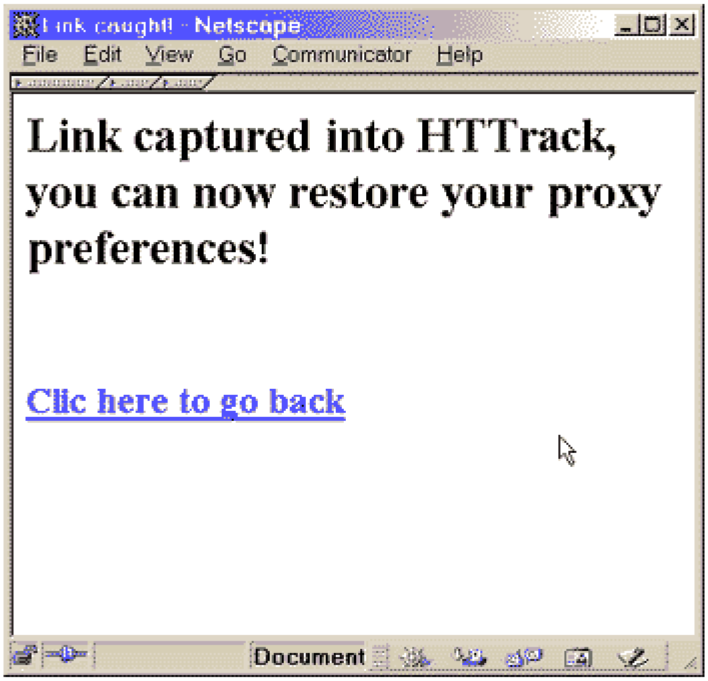


Please temporarily set your proxy preferences in your browser to the following values (cut/paste proxy address and port).
Then, in your browser page, click on the submit button of the form, or click on the specific link you want to capture.

Please TEMPORARILY set your browser's proxy preferences to:

Proxy's address:
192.168.0.1
Proxy's port:
8080

This will capture the desired link from your browser to HTTrack.



Link captured into HTTrack, you can now restore your proxy preferences!

Clic here to go back

5. Click on the **NEXT** button

6. Go to the next step...

Step 3 : Ready to start

1. If you want, you may connect immediately or delay the mirror

   If you don't select anything, HTTrack will assume that you are already connected to the Internet and that you want to start the mirror action now

   - Connect to this provider

     You can select here a specific provider to connect to when beginning the mirror if you are not already connected to the Internet.

   - Disconnect when finished

     Click on this checkbox to ask httrack to disconnect the network when mirror is finished.

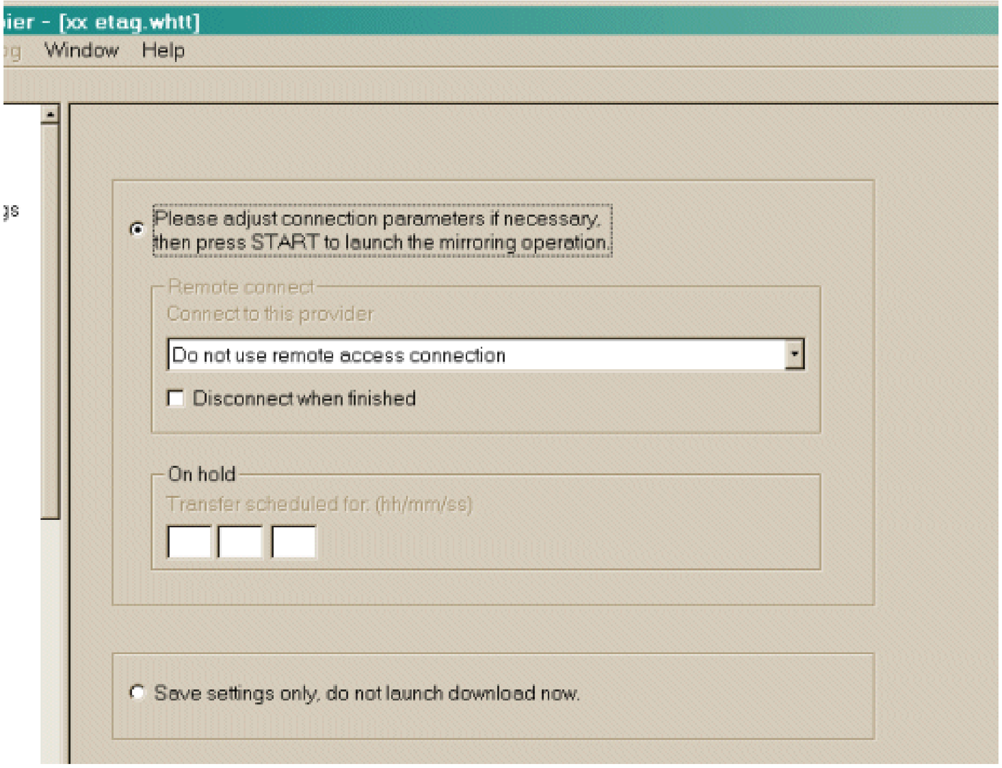
   - Shutdown PC when finished

     Click on this checkbox to ask httrack to shutdown your computer when mirror is finished.

   - On Hold

     You can enter here the time of the mirror start. You can delay up to 24 hours a mirror using this feature.

2. Click on the **FINISH** button



● Please adjust connection parameters if necessary, then press START to launch the mirroring operation.

Connect to this provider: Do not use remote access connection

□ Disconnect when finished

Transfer scheduled for: (hh/mm/ss)

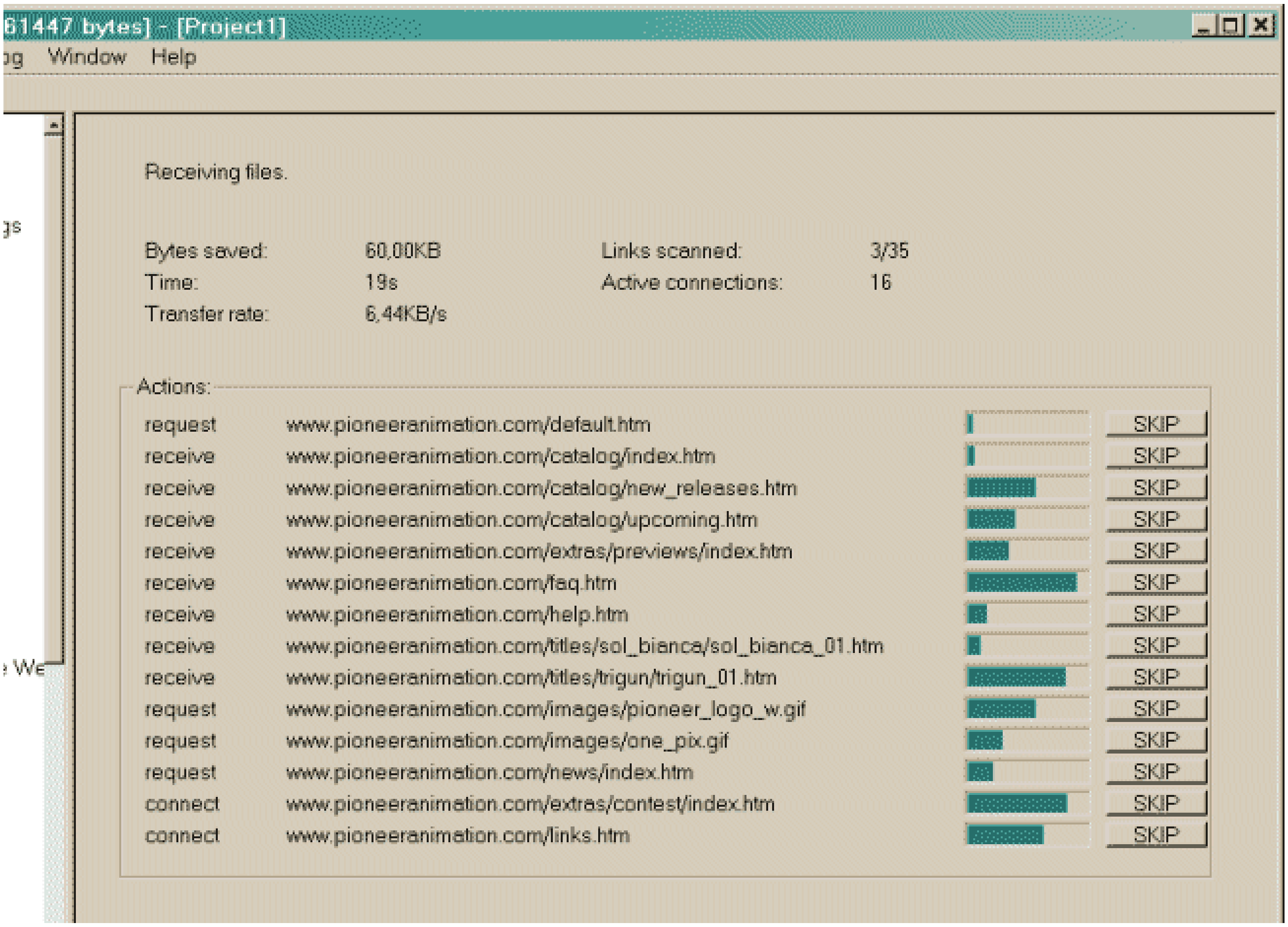
○ Save settings only, do not launch download now.

Step 4 : Wait!

1. Wait until the mirror is finishing

   You can cancel at any time the mirror, or cancel files currently downloaded for any reasons (file too big, for example)

   Options can be changed during the mirror: maximum number of connections, limits...



Receiving files.

Bytes saved: 60,00KB
Time: 19s
Transfer rate: 6,44KB/s
Links scanned: 3/35
Active connections: 16

Actions:

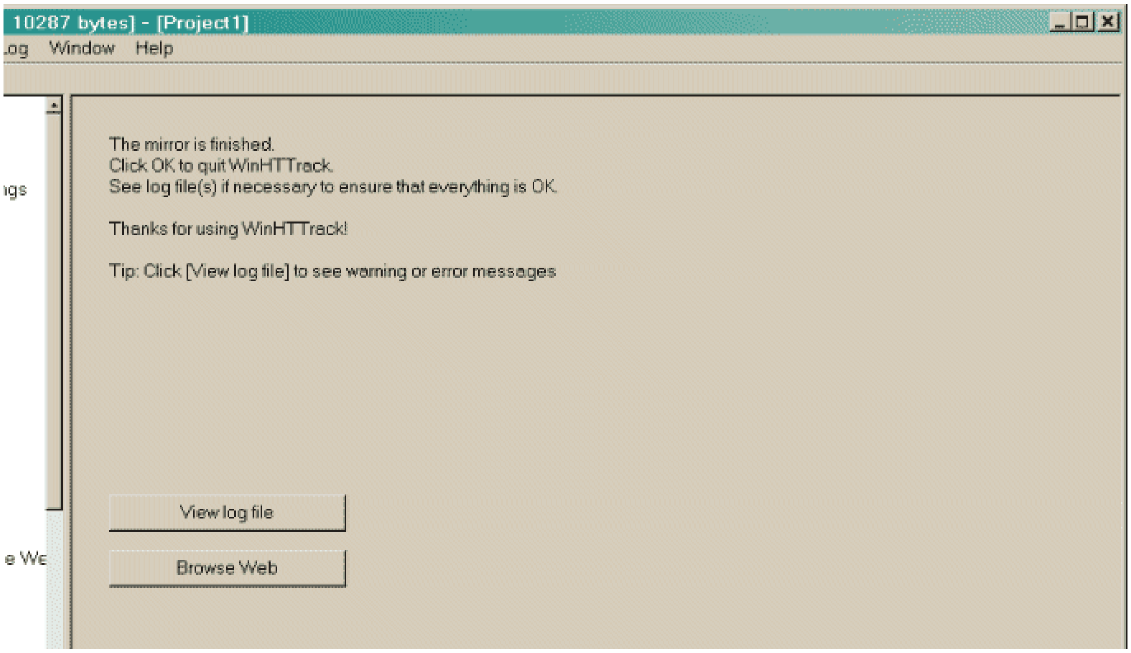
request	www.pioneeranimation.com/default.htm	SKIP
receive	www.pioneeranimation.com/catalog/index.htm	SKIP
receive	www.pioneeranimation.com/catalog/new_releases.htm	SKIP
receive	www.pioneeranimation.com/catalog/upcoming.htm	SKIP
receive	www.pioneeranimation.com/extras/previews/index.htm	SKIP
receive	www.pioneeranimation.com/faq.htm	SKIP
receive	www.pioneeranimation.com/help.htm	SKIP
receive	www.pioneeranimation.com/titles/sol_bianca/sol_bianca_01.htm	SKIP
receive	www.pioneeranimation.com/titles/trigun/trigun_01.htm	SKIP
request	www.pioneeranimation.com/images/pioneer_logo_w.gif	SKIP
request	www.pioneeranimation.com/images/one_pix.gif	SKIP
request	www.pioneeranimation.com/news/index.htm	SKIP
connect	www.pioneeranimation.com/extras/contest/index.htm	SKIP
connect	www.pioneeranimation.com/links.htm	SKIP

2. Go to the next step...

Step 5 : Check the result

1. Check log files

   You may check the error log file, which could contain useful information if errors have occurred



The mirror is finished.
Click OK to quit WinHTTrack.
See log file(s) if necessary to ensure that everything is OK.

Thanks for using WinHTTrack!

Tip: Click [View log file] to see warning or error messages

2. See the troubleshooting page