WP5: Tool tests for scenario 2.4

Authors:

Eva Toller (National Archives of Sweden, RA)
……

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	**Public**	**P**
C	**Confidential, only for members of the consortium and the Commission Services**	

Revision History

Revision	Date	Author	Organisation	Description
0.1	20130618	Eva Toller	RA	First draft
0.2	20130619	Eva Toller	RA	Added partial tests results for WARC Tools (moved to PoC #2).
0.3	20130620	Eva Toller	RA	Added partial tests results for Web Curator Tool
0.4	20130624	Eva Toller	RA	Added partial tests results for SWAT
0.5	20130730	Eva Toller	RA	Added results for HTTRack
0.6	20130805	Eva Toller	RA	Added more partial tests results for SWAT
0.7	20130806	Eva Toller	RA	Added more results for Web Curator Tool
0.8	20130808	Eva Toller	RA	Added results for Heritrix (not installed)

1 TOOL TESTING FOR SCENARIO 2.4

1.1 SCENARIO DESCRIPTION

“A history student interested in natural history discovers that Riksarkivet has archived the "Linnéjubilet" web site http://www.riksarkivet.se/default.aspx?id=23153 .He wonders how he can get access to it (the link www.linne2007.se obviously doesn't work anymore).“

General comment: actually, the link www.linne2007.se *does* still work, but for test purposes this is not important. Furthermore, there can be no guarantee that it will continue to be accessible.

Suggested test data: see document DCH-RP_WP5_Scen-2-4_ID-68.pdf

1.2 DISPOSITION

Chapter 2 and the following chapters are structured in the following way:

In sections X.1, a short description is given of the tool and how it works.

In sections X.2, the data set(s) that the tool will be tested on is described. If there are several data sets, they are described in sub sections: X.2.1, X.2.2, X.2.3 … X.2.n.

In sections X.3, the results of the tests are given (if any). If there are several data sets and the results differ significantly between them, they are described in sub-sections: X.3.1, X.3.2, X.3.3 … X.3.n.

In sections X.4, general comments are given about the tool and its usability for digital cultural heritage preservation, dissemination et c. (This section may be skipped if it was not possible to install and/or run the tool).

1.3 TEST ENVIRONMENT

When nothing else is said, the test environment is a PC (Personal Computer) with Windows 7 Professional, processor Intel(R) 2,7 GHz, and 8 GB working memory (RAM).

2 HTTRACK

2.1 GENERAL DESCRIPTION

“HTTrack is a free (GPL, libre/free software) and easy-to-use offline browser utility. It allows you to download a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to your computer.”

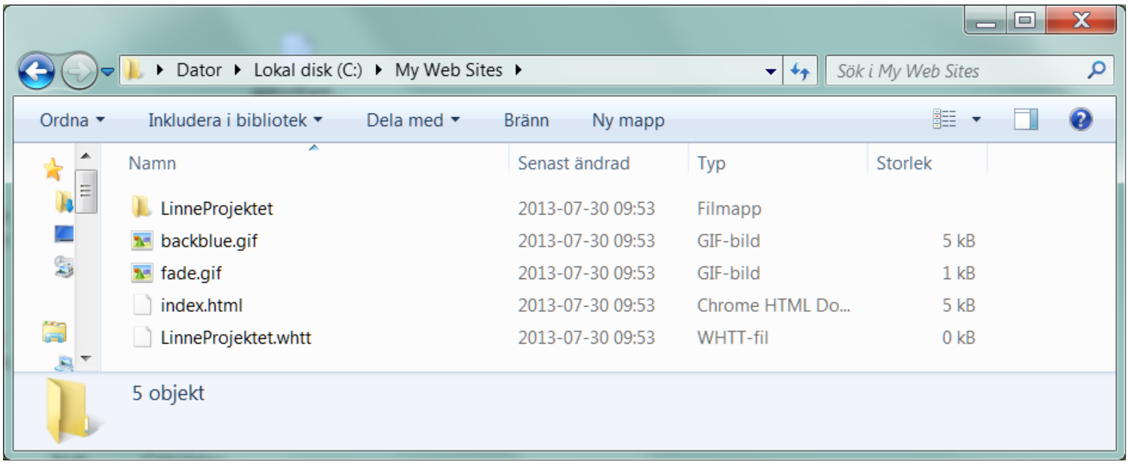
2.2 DATA SET

The web site “Linnéjubiléet” (Linné Anniversary) was constructed for the 300th anniversary of the birth of Carl von Linné. It contains information about specific anniversary activities, about Linné and his life, about gardens and exhibitions, and much more. The site consists of 4058 files altogether, and the total size is 469 MegaByte.

For details, see: **DCH-RP_WP5_Scen-2-4_ID-68.pdf**

2.3 TEST RESULTS

The quality of the downloaded web site (stored on **C:\My Web Sites**) seems, in the first examination, to be as good as the original web site. One drawback is that you may get a little confused when trying to find how to open the downloaded version. After the download, the directory **C:\My Web Sites** has the following contents:



Namn	Senast ändrad	Typ	Storlek
LinneProjektet	2013-07-30 09:53	Filmapp	
backblue.gif	2013-07-30 09:53	GIF-bild	5 kB
fade.gif	2013-07-30 09:53	GIF-bild	1 kB
index.html	2013-07-30 09:53	Chrome HTML Do...	5 kB
LinneProjektet.whtt	2013-07-30 09:53	WHTT-fil	0 kB

5 objekt

Figure 1: Contents of C:\My Web Sites

There are (at least) three ways to open the downloaded web site:

1. Choose “LinneProjektet”, then “index.html”. **or**
2. Choose “LinneProjektet”, then “www.linne2007.se”, then “index.html”. **or**
3. Choose “index.html”, then (from the web page that opens), choose “LinneProjektet”.

2.4 USABILITY

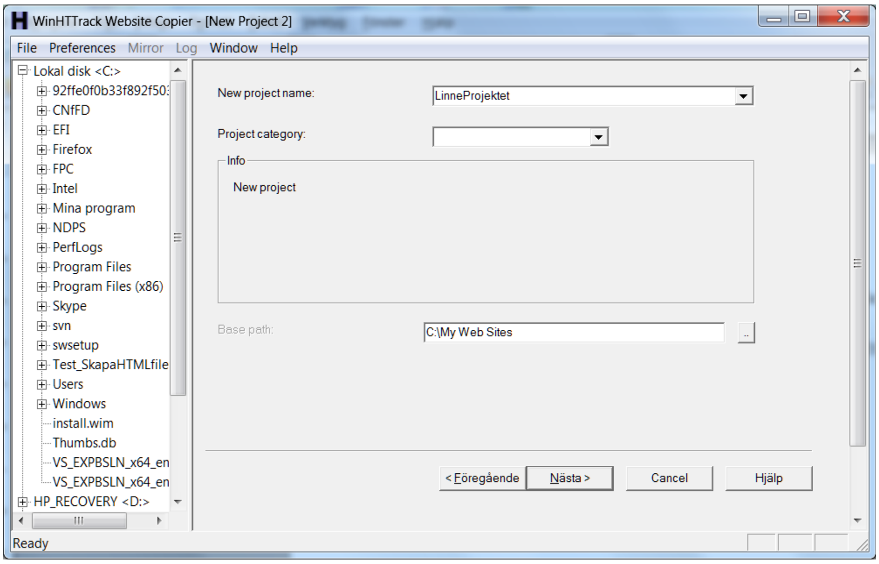
2.4.1 Download and installation

The downloadable file is a simple .exe file: **httrack-3.47.21.exe**.

The usual pop-up windows that can be expected during program installations are shown, for example:

- Acceptance of Licence Agreement
- Selection of the directory for installation
- Where to place shortcut folder

When you start HTTrack the first time, a window about language preference pops up. Then you can change the language from the default (English), to (for example) Swedish. To implement the change, you have to exit and restart HTTrack. The built-in manual is in English. (However, although I did not choose to change the language from English, there are still some text in Swedish here and where, especially button labels).



New project name: LinneProjektet

Project category:

Info

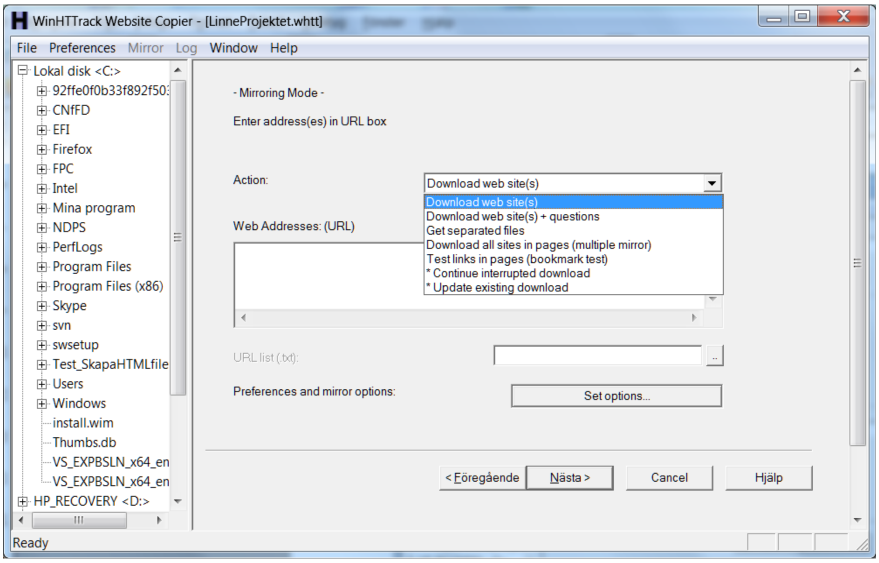
New project

Base path: C:\My Web Sites

Figure 2: Start page for HTTrack

After clicking on “Nästa” (= “Next”), the first thing you must do is to start a new project (which I named “LinneProjektet”). Then you can fill in a project category (however, there is no information about what categories there are to choose from). However, it is possibly to go forward (clickin on “Nästa” again) without supplying a project category. Before you continue, you may also choose a root path for the project (default is “C:\My Web Sites”).

Now, you can start to download a website. You can choose the default option "Download web site(s)" but you can also modify the download by choosing other options (see the figure below).


WinHTTrack Website Copier - [LinneProjektet.whtt]

- Mirroring Mode -

Enter address(es) in URL box

Action: Download web site(s)

- 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)

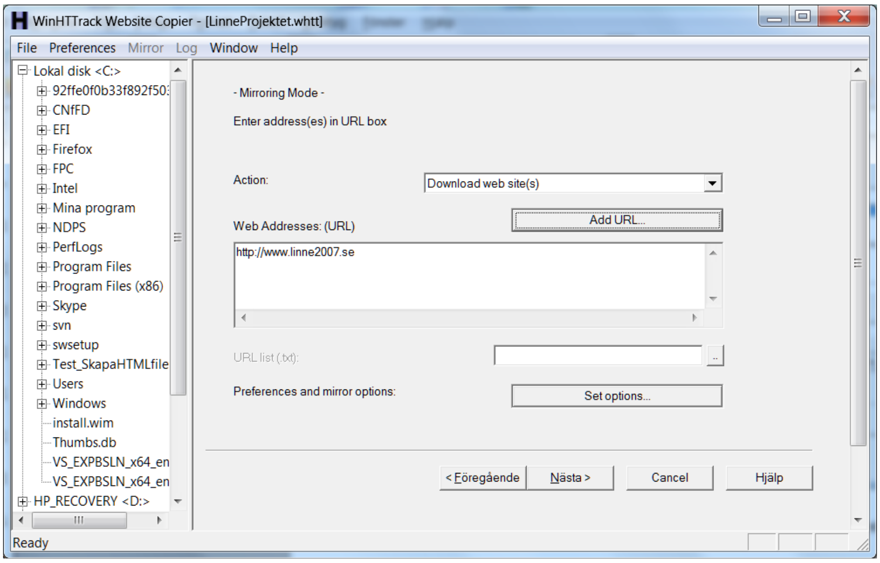
URL list (.txt):

Preferences and mirror options:

Ready

Figure 3: Choosing options for downloading a web site

When you click on "Add URL" (a button hidden by the option list in Figure 2) you can write www.linne2007.se. The, the URL is added to the Web Addresses area:


WinHTTrack Website Copier - [LinneProjektet.whtt]

- Mirroring Mode -

Enter address(es) in URL box

Action: Download web site(s)

Web Addresses: (URL)

http://www.linne2007.se

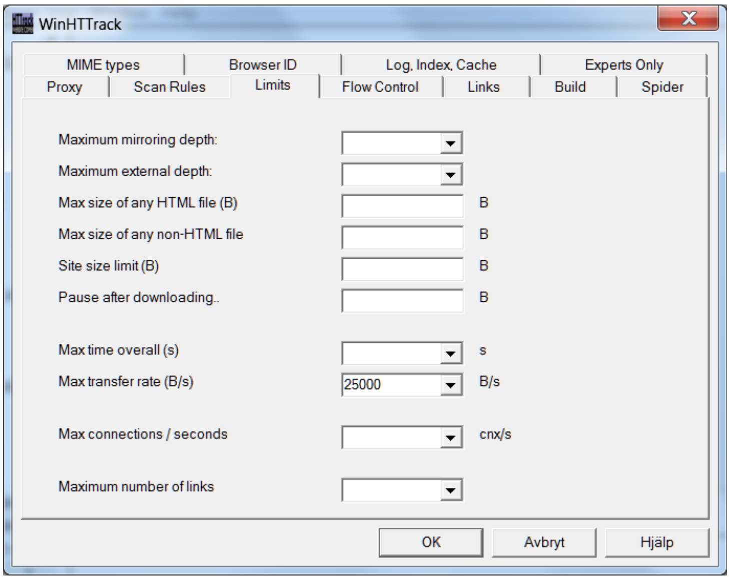
URL list (.txt):

Preferences and mirror options:

Ready

Figure 4: Choosing an URL for downloading

If "Set options" is clicked, you can choose a number of other parameters (for example, what MIME identity different file types should correspond to).



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) 25000 B/s

Max connections / seconds cnx/s

Maximum number of links

Figure 5: Options for modifying the download

I did choose to *not* modify any of these options.

There are some last things you can modify about the downloading process itself (for example, to disconnect when finished). Then you click on the button “Slutför” (= “Finish”).

The actual download operation only took a couple of seconds. When completed, you can choose to view the log file and/or browse the mirrored web site (that is, the one you just have downloaded).

The log file is short and contains the following information for this download:

```
HTTrack3.47-21+htsswf+htsjava launched on Tue, 30 Jul 2013 09:53:37 at http://www.linne2007.se +*.png
+*.gif +*.jpg +*.css +*.js -ad.doubleclick.net/* -mime:application/foobar

(winhttrack -qwC2%Ps2u1%s%uN0%I0p3DaK0H0%kf2A25000%f#f -F "Mozilla/4.5 (compatible; HTTrack 3.0x; Windows
98)" -%F "<!-- Mirrored from %s%s by HTTrack Website Copier/3.x [XR&CO'2013], %s -->" -%l "en, en, *"
http://www.linne2007.se -O1 "C:\My Web Sites\LinneProjektet" +*.png +*.gif +*.jpg +*.css +*.js
-ad.doubleclick.net/* -mime:application/foobar )

Information, Warnings and Errors reported for this mirror:

note: the hts-log.txt file, and hts-cache folder, may contain sensitive information, such as
username/password authentication for websites mirrored in this project

 do not share these files/folders if you want these information to remain private

HTTrack Website Copier/3.47-21 mirror complete in 2 seconds : 2 links scanned, 1 files written (210 bytes
overall) [2309 bytes received at 1154 bytes/sec]

(No errors, 0 warnings, 0 messages)
```

Thus, the download was completed without errors.

2.4.2 Recommendation

Since it was both easy to install and use this tool, and the quality of the result also was good, this should be a suitable tool for the downloading of web sites when the most important aim to give easy access to end users.

However, it remains to be investigated how good the downloaded format is for long-term preservation, and also how efficient the program is when many web sites are downloaded as a batch, simultaneously. For preservation, it would be useful to test all the different options that you can set for a download.

Grade

On a scale from 1 (very bad) to 5 (very good). X is “Not applicable”.

Simplicity of installation: 5

Simplicity of management: X

Ease of use: 4 – 5

Generality of solution: 5

Quality of result: 5

3 SWAT (SNAPPY WEB ARCHIVING TOOL)

3.1 GENERAL DESCRIPTION

“SWAT (Snappy Web Archiving Tool) is a tool designed for archiving web sites and displaying the archive in a simple way. Besides harvesting all files from the web site, SWAT generates snapshots of each page to TIFF files and describes the entire archive in a METS-file.”

3.2 DATA SET

The data set that was to be used was the same as for HTTrack (see section 2.2).

3.3 TEST RESULTS

Not applicable.

3.4 USABILITY

3.4.1 Download and installation

On the web page http://sourceforge.net/projects/swat-archiving/ , there is a very brief description of SWAT. There is also a button for downloading.

A longer description on SWAT is provided on http://swat-archiving.sourceforge.net/ . There, the several necessary steps are described that you have to go through to create the web archive (“webchive”).

The download results in a file **swat-0.6.tar.bz2**.

http://7-zip.org/

The **.tar.bz2** can be unpacked by the free tool 7-zip (http://www.7-zip.org/ . It has to be done in two steps: first, the file is unpacked *to* a **.tar** file, then you unpack the **.tar** file.

The developer of SWAT provided a link to a report about SWAP, including installation instructions (section 6.8): http://swat-archiving.svn.sourceforge.net/viewvc/swat-archiving/doc/long_term_digital_preservation_of_web_sites.pdf?revision=6

According to the developer of SWAT, it can be run on Windows but is easiest to run on a Linux-based system (for example, Ubuntu). But it should also be possible to run it on Windows.

According to the manual, the following software should be installed:

1. Ruby (http://www.ruby-lang.org/)
2. Mongrel (http://mongrel.rubyforge.org/)
3. Java (http://www.java.com/)

4. DROID (http://droid.sourceforge.net/)
5. QT (http://qt.digia.com/). This is only an evaluation license (30 days).
6. CutyCapt (http://cutycapt.sourceforge.net/)
7. ImageMagick (http://www.imagemagick.org/)
8. GNU Wget (http://www.gnu.org/software/wget/)
9. SWAT (http://swat-archiving.sourceforge.net/)
10. A database server, the author recommends SQLite (http://www.sqlite.org/)

All tools could be downloaded and unpacked. However, when Ruby was to be downloaded, the downloader recommended by http://www.ruby-lang.org/ (RubyInstaller) was blocked by F-Secure, because this site has been reported as containing malware (malicious software, code infected by viruses or trojans) and should be avoided.

The other choices you had was to download and compile the source code yourself, or using third party tools. I actually managed to download an executable version from the RubyInstaller site (from the archive) but that is *not* an action to be recommended.

This, together with the many tools that you have to download, unpack, and install, makes it doubtful if this tool can and/or should be managed by small institutions, that have only one or a few web sites to preserve and present for end-user access. This is also the case with Scenario 2.4; “Linnéjubileet” was a small temporary government agency with only one web site.

On a scale from 1 (very bad) to 5 (very good):

Simplicity of installation: 2 ???

4 WARC TOOLS

4.1 GENERAL DESCRIPTION

“The main goal of WARC Tools is to facilitate and promote the adoption of the WARC file format for storing web archives by the mainstream web development community by providing an open source software library, a set of command line tools, web server plug-ins and technical documentation for manipulation and management of web archive files, or WARC files. WARC files are produced by web archiving crawlers, such as Heritrix, the open-source, extensible, Web-scale, archiving quality Web crawler developed by the Internet Archive with the Nordic National Libraries, and Hanzo's own commercial crawlers.”

4.2 DATA SET

The data set that was to be used was the same as for HTTrack (see section 2.2).

4.3 TEST RESULTS

Not applicable.

4.4 USABILITY

4.4.1 Download and installation

WARC Tools are recommended by the **LDB-centrum** (LDP Centre, Centre for Long-term Digital Preservation, see http://www.ltu.se/centres/Centrum-for-langsiktigt-digitalt-bevarande-LDB?l=en).

However, the page that is referred to from http://www.ltu.se/centres/Centrum-for-langsiktigt-digitalt-bevarande-LDB/Bibliotek/Webb/Om-filformatet-WARC-och-verktyg-for-formatet-1.67311 (in Swedish), https://code.google.com/p/warc-tools/, is now obsolete:

> “All development has ceased on this version of warctools.
>
> A newer version in python has been published and is maintained on our site here: http://code.hanzoarchives.com/warc-tools“

There are no downloadable executable software files any more at https://code.google.com/p/warc-tools/.

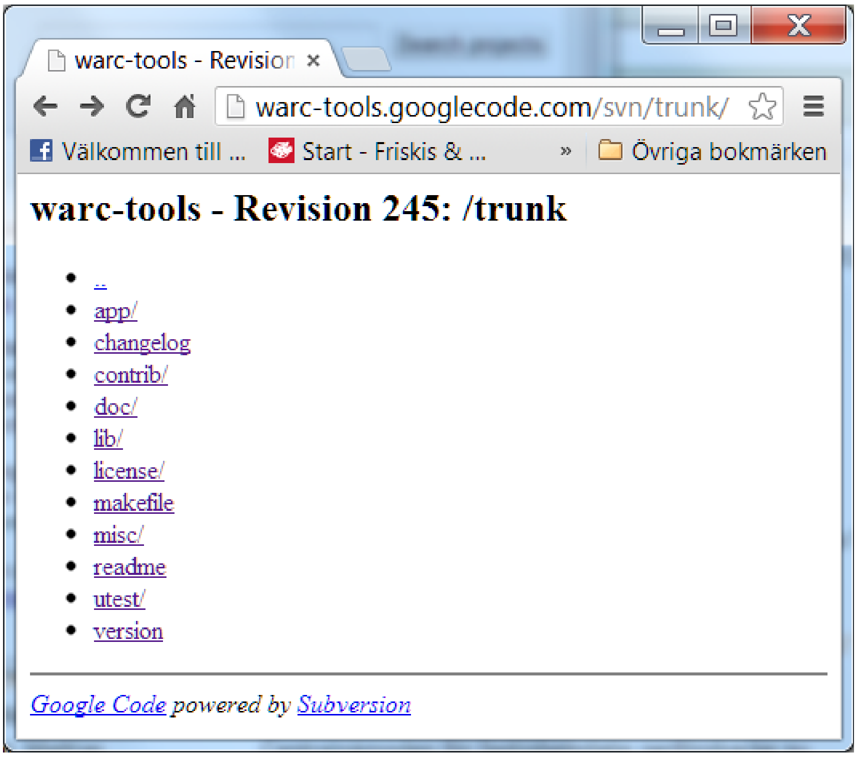
On http://code.hanzoarchives.com/warc-tools , there are no immediately recognizable files for download.

In the “Downloads” page (http://code.hanzoarchives.com/warc-tools/downloads), the message “There are no files available to download.” is shown.

On the “Source” page (https://code.google.com/p/warc-tools/source/checkout) there is source code that can be obtained:

http://warc-tools.googlecode.com/svn/trunk/ warc-tools-read-only

When accessing http://warc-tools.googlecode.com/svn/trunk/ , there is no item “warc-tools-read-only” on the top level:



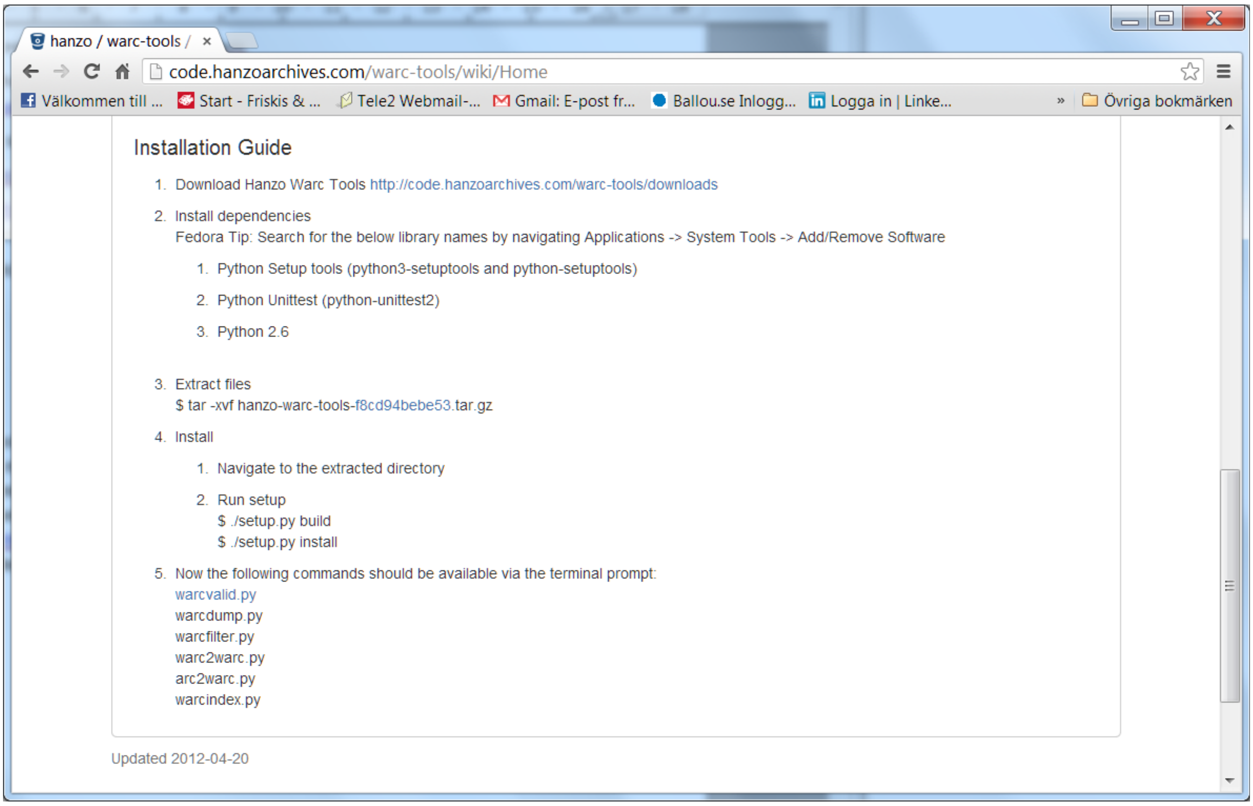
warc-tools - Revision 245: /trunk

- ..
- app/
- changelog
- contrib/
- doc/
- lib/
- license/
- makefile
- misc/
- readme
- utest/
- version

Google Code powered by Subversion

Figure 6: Contents of http://warc-tools.googlecode.com/svn/trunk/

On the “Wiki for WARC Tools” page (http://code.hanzoarchives.com/warc-tools/wiki/Home), there are the following installation instructions:



Installation Guide

1. Download Hanzo Warc Tools http://code.hanzoarchives.com/warc-tools/downloads
2. Install dependencies
   Fedora Tip: Search for the below library names by navigating Applications -> System Tools -> Add/Remove Software
   1. Python Setup tools (python3-setuptools and python-setuptools)
   2. Python Unittest (python-unittest2)
   3. Python 2.6
3. Extract files
   $ tar -xvf hanzo-warc-tools-f8cd94bebe53.tar.gz
4. Install
   1. Navigate to the extracted directory
   2. Run setup
      $ ./setup.py build
      $ ./setup.py install
5. Now the following commands should be available via the terminal prompt:
   warcvalid.py
   warcdump.py
   warcfilter.py
   warc2warc.py
   arc2warc.py
   warcindex.py

Updated 2012-04-20

Figure 7: Instructions for installation

The first link leads to the already visited “Downloads” page (http://code.hanzoarchives.com/warc-tools/downloads), where the message “There are no files available to download.” was shown.

It is also obvious that you must install Python. Exactly how to do this is unclear.

On a scale from 1 (very bad) to 5 (very good):

Simplicity of installation: 1

5 WEB CURATOR TOOL

5.1 GENERAL DESCRIPTION

“The Web Curator Tool (WCT) is an open-source workflow management application for selective web archiving. It is designed for use in libraries and other collecting organisations, and supports collection by non-technical users while still allowing complete control of the web harvesting process. It is integrated with the **Heritrix** web crawler and supports key processes such as permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata.”

5.2 DATA SET

The data set that was to be used was the same as for HTTrack (see section 2.2).

5.3 TEST RESULTS

Not applicable.

5.4 USABILITY

5.4.1 Download and installation

On the web page http://webcurator.sourceforge.net/ , which is quite comprehensible, you can download a User Manual, Release Notes, and a zip file with the code.

You can choose between downloading a **.zip** file and a **.tar.gz** file.There is also a readme.txt file (it contains the same information as the Release Notes). Unfortunately, the User Manual does not contain any instructions for installation.

However, if you instead look among the documents for Developers, there is a guide for System Administrator that seems to contain all the necessary information. (Note that it is *not* intuitive that such a manual is to be found among information for Developers).

The installation instructions for WCT and the other necessary software are written for Windows 2003, so all of them may not be applicable for Windows 7.

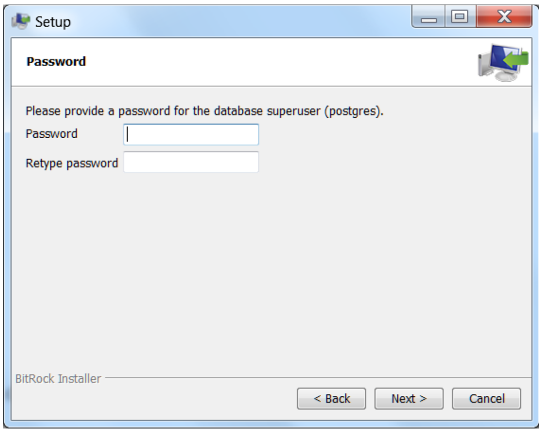
The downloaded **.zip** file is contains the following directories:

- **docs** (a lot of documentation, mostly in PDF format)
- **etc** (two **.jar** files)
- **sql** (a lot of **.sql** files for Oracle, PostGres, and MySQL)
- **upgrade** (a lot of **.sql** files for Oracle, PostGres, and MySQL)
- **war** (three **.war** files)

You have to install one of the database systems Oracle, PostGreSQL, or MySQL to make WCT work. **PostGreSQL** is chosen.

PostGreSQL can be downloaded from http://www.postgresql.org/download/ . There are different options (for example, you can download the source code), but the simplest way is do download the binary code for Windows (a single .exe-file).

The installation of PostGreSQL is rather easy (and you don't have to create an account, as you nowadays have to do if you choose to install MySQL). You have to supply a password for the “superuser” **postgres**:



Password

Please provide a password for the database superuser (postgres).

Password

Retype password

Figure 8: Supply a password

You also have to select a port number (5432 is default). After that, the installation was automatic.

The WCT instructions for installing PostgreSQL are not the same as for the actual PostgreSQL installation, but much more complicated (and probably wholly outdated).

Next, WCT itself should be downloaded and extracted. **Java SE Development KIT** is also necessary.

Another installation that must be made: **Tomcat.** Here, you must also change a lot of environment variables for Java. You also have to install jar files.

Finally, you have to deploy WCT, and also make a lot of configurations.

The final words in the installation section (which is six pages long) are the following:

> *“It is clear, that this is not Windows based application, which does not mean it doesn't run as well as it does on Linux, it only means that it needs a bit of tweaking and configuring before you can actually run it.* ***We will try, with the help of WCT development team, to give you updated information on how to deploy newest versions of WCT on Windows.”***

Then follows six more pages with lists of different settings you should do.

There seems to be hard or impossible to find some current installation instructions. This, together with the many tools that you have to download, unpack, and install, makes it doubtful if this tool can and/or should be managed by small institutions, that have only one or a few web sites to preserve and present for end-user access. This is also the case with Scenario 2.4; “Linnéjubileet” was a small temporary government agency with only one web site.

On a scale from 1 (very bad) to 5 (very good):

Simplicity of installation: 1

6 HERITRIX

6.1 GENERAL DESCRIPTION

“Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project.”

6.2 DATA SET

The data set that was to be used was the same as for HTTrack (see section 2.2).

6.3 TEST RESULTS

Not applicable.

6.4 USABILITY

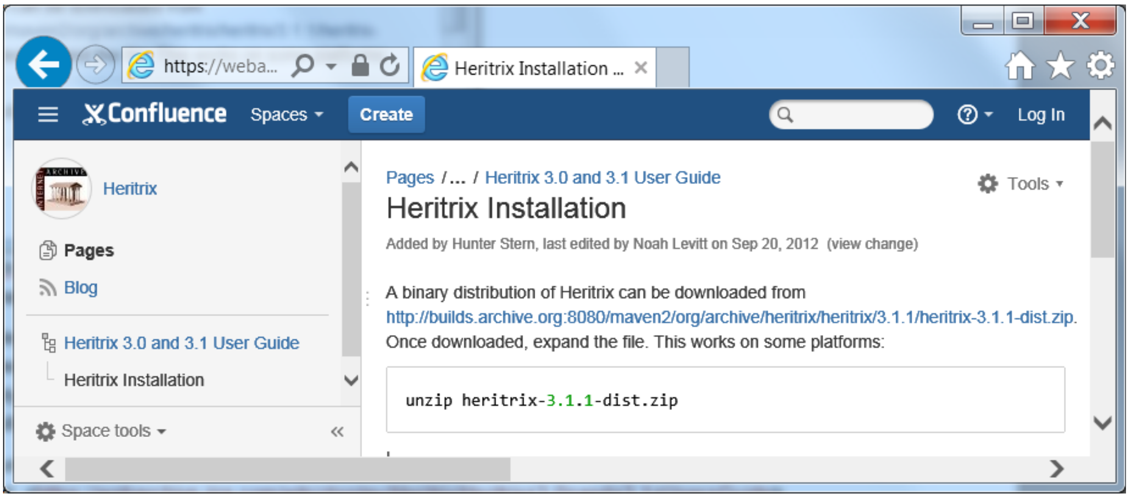
6.4.1 Download and installation

On the web page https://webarchive.jira.com/wiki/display/Heritrix/Heritrix , you can go directly to Downloads and get either a **.tar.gz** file or a **.zip** file. It is not stated in the start page what platforms that Heritrix can be used on. When searching with the phrase “heritrix platforms”, I found the following information (http://crawler.archive.org/articles/user_manual/install.html):

> “Because Heritrix is a pure Java program it can (in theory anyway) be run on any platform that has a Java 5.0 VM. However we are only committed to supporting its operation on Linux and so this chapter only covers setup on that platform. Because of this, what follows assumes basic Linux administration skills. Other chapters in the user manual are platform agnostic.“

After unpacking the download, you can read in the readme.txt file that there is a User Manual for Getting started: <https://webarchive.jira.com/wiki/display/Heritrix/Heritrix+3.0+and+3.1+User+Guide>

A ctrl-click will take you there. On this page, there is a headline “Heritrix installation”. However, this installation page only contains the following information:



Pages / ... / Heritrix 3.0 and 3.1 User Guide

Heritrix Installation

Added by Hunter Stern, last edited by Noah Levitt on Sep 20, 2012 (view change)

A binary distribution of Heritrix can be downloaded from http://builds.archive.org:8080/maven2/org/archive/heritrix/heritrix/3.1.1/heritrix-3.1.1-dist.zip. Once downloaded, expand the file. This works on some platforms:

```
unzip heritrix-3.1.1-dist.zip
```

Figure 9: “Installation instructions”

http://crawler.archive.org/articles/user_manual/install.html actually contains a little more information, but it is not so easy to find. If you go to “Documentation” from https://webarchive.jira.com/wiki/display/Heritrix/Heritrix , the first thing you see is the previously mentioned page with the “installation instructions”. The second thing you see is javadoc, obviously only for developers. There is also a user manual: http://crawler.archive.org/articles/user_manual/index.html . It contains a little more information about installation, but only in the form of command-line interface text.

It is unclear if the tool can run on other platforms than Linux. According to the FAQ, this has been tried even if it is not supported.

A benefit with this tool is that a lot of third-party products don't seem to be required (besides Linux, only Java Runtime Environment is mentioned). However, the installation instructions are not sufficient for an inexperienced user.

On a scale from 1 (very bad) to 5 (very good):

Simplicity of installation: 1 – 2