Tests of software and strategies for micro-archiving websites

Bo Hovgaard Thomasen, MA student,
Department of Information and Media Studies,
University of Aarhus, e-mail: bht@imv.au.dk

Archiving websites or parts of websites for use in research or study projects means that a number of theoretical, methodical, pragmatic, and technical questions must be thought through and clarified. With the book, *Archiving Websites. General Considerations and Strategies*[1], Niels Brügger is concerned with the theoretical and methodical aspects of archiving websites for use by researchers or students.[2] Based on this work, a test of various types of archiving software leading to web pages archived in varying formats has been carried out. Thus, we are looking at the archiving of complete websites or individual web pages in a static form, as screenshots, or as screen recordings. These four methods of archiving are discussed in depth in the above-mentioned book, while the present text will attempt to supplement Brügger's thoughts, by giving an account of tests of specific archiving software carried out in order to contribute pragmatic and technical experience from the operationalising of the book's theoretical and methodical deliberations.

At the same time, this text aims to present the considerations and choices made, with the aim of increasing the validity of the test. A further goal is to allow those who wish to verify the test or to test other archiving programmes to do so based on this explication.

The main questions are what specific computer programmes can be used for the micro-archiving of websites, and what are the advantages, limitations, and deficiencies to be found in archiving with the programmes tested. The tests

[1] The book is one in a series of books from the Centre for Internet Research, URL: http://cfi.imv.au.dk/pub/boeger/bruegger_archiving.pdf

[2] Brügger differentiates between macro- and micro-archiving of websites. Macro-archiving covers the archiving of many websites on a large scale, for instance with the aim of preserving the digital cultural heritage. Micro-archiving is the archiving of websites on a smaller scale, usually for a definite purpose, such as archiving the empiricism of a research project. (Brügger 2005: 9).

have been carried out by MA student Bo Hovgaard Thomasen from July-December 2004.

Selecting software for testing

The first step in fulfilling the aim of verifying the archiving strategies Brügger puts forth in his book was to identify the computer programmes to be included in the test. The programmes in the test have been chosen on the basis of: (1) their functionality, (2) their availability to researchers and students, as regards both procurement and use.

Thus, an effort has been made to find programmes for the test that could either be used free of charge, or where as a minimum, fully functional 30-day trial versions were available. In addition, in order to ensure the broad span of the test, it was a criterion that programmes for both the Windows (98, ME, NT 4, 2000 and XP) and Mac OS X (from version 10.3) platforms were included in the test. One programme "independent of platform" was also chosen, written in Java. The first reason for this criterion is the desire to test the most relevant programmes, without excluding any in advance through a one-sided focus on one operative system. The second reason is that the test hopes to apply to all researchers and students, not just those who work with Windows or Mac OS, respectively. Several of the programmes selected can also be executed on UNIX systems, such as Linux, which further increases the test's span. However, for practical reasons it was necessary to limit the test to the programmes selected for Windows and Mac OS X. Therefore technical directions for the execution of the programmes in other operative systems than the two above-mentioned have not been devised.

The programmes were selected after an initial pilot study. In order to include the most relevant programmes in the test, a systematic search was made of various Internet archives as well as the Google.com search engine. The most relevant programmes for archiving a website are those that comply with the archiving of the forms of expression used for the construction of the web page at any given time. These forms of expression are constantly being further developed, so that in choosing programmes for the test an effort was made to ensure that the programmes were as new as possible; it must be assumed that the newest technologies in the construction of web pages can best be archived by using the latest archiving programmes rather than older versions. The attempt to comply with this was made by searching software archives, which have the advantage of being updated as soon as new programmes are published. In addition, the Google.com search engine was used to supplement the archive search, in order to include software that for one reason or another is not available in the software archives. This includes certain commercial programmes as well as programmes that are only distributed via the programme's web page. Table 1 shows the search sites and search keys used in the searches.

Table 1: Search sites and search keys

Search sites		**Search key**
Download.com	http://www.download.com	offline browser, offline browsing, screen film, save website (/web page), web archiving, copy website (/web page), copier website (/web page), web snapshot, web screen dump
Google	http://www.google.com	
Macupdate	http://www.macupdate.com	
Snapfiles	http://www.snapfiles.com	
Tucows	http://www.tucows.dk	
Versiontracker	http://www.versiontracker.com	

The available resources for the test allowed for the selection of a total of eighteen programmes for testing. The searches and a subsequent provisional test resulted in a choice of the eighteen most promising programmes. Table two shows the programmes tested.

Test of archiving software

Two different tests were carried out on each archiving programme. First a test of functionality, where the programme's ability to archive the various elements of which a website is composed was tested. This part of the test also had the goal of determining recommended settings and parameters that are the most advantageous when using the programme in archiving. Secondly, this was followed up with a test of the programme's speed, using the settings determined as recommendations in the test of functionality. The speed test resulted in an assessment of how quickly the programme is able to archive a website, how much hard-disk space the archived material requires, and to what extent the presence of a person is required during the archiving process. These variables are not all relevant to all the archiving modes tested, and have therefore been omitted in some cases. For instance, archiving speed is not relevant when archiving a screenshot, as a screen shot is archived "instantaneously." Similarly, screen recording is highly dependent on the person doing the recording: how quickly they can navigate through the desired web pages, test the desired elements on the web page, or record a streamed film cut. The individual test results show whether the speed test parameter has been used for the programme in question.

Table 2: Archiving programmes included in the test.

Programme	**Vers.**	**Price**	**OS**	**Download**	**Notes**
Adobe Acrobat Professional	6.01	5187,50 DKK	Win-dows/ Mac OS X	http://adobe.dk/products/acrobatpro/main.html	*Complete website* 30-day trial version for Windows can be downloaded.
DeepVacuum	1.24	7,00 USD	Mac OS X	http://www.hexcat.com/deepvacuum/	*Complete website* Mac OS X programme, based on the command-line programme 'wget'.
The utility Save As PDF...	-	Inherent to Mac OS X	Mac OS X	-	*Individual web page in a static state*
The utility Print Screen	-	Inherent to Windows	Windows	-	*Screenshot*
Grab	1.2	Accompanies Mac OS X	Mac OS X	-	*Screenshot*
Microsoft Internet Explorer	6.0	Free	Windows	http://www.microsoft.dk	*Complete website*
Microsoft Internet Explorer	5.2.3	Free	Mac OS X	http://www.microsoft.dk/mac	*Complete website*
MM3-WebAssistant Private	2005	Free	Java	http://www.mm3tools.de/WebAssistant	*Complete website* Saves web pages visited by the browser in a cache for offline use.
Paparazzi!	0.1.8	Free	Mac OS X	http://0x.se/paparazzi/	*Individual web page in a static state*
PrimoPDF	1.0	Free	Windows	http://www.primopdf.com	*Individual web page in a static state*
SnagIt	7.1.2	39,95 EUR	Windows	http://www.techsmith.com/products/snagit/	*Individual web page in a static state, screenshot, screen recording* 30-day trial version can be downloaded
Snapz Pro X	2.0	69,00 USD	Mac OS X	www.ambrosiasw.com/utilities/snapzprox/	*Screenshot, screen recording* 30-day trial version can be downloaded
Web2Pic	1.1	Free	Windows	http://www.anloer.com/web2pic/	*Individual web page in a static state*
Webkit2png	0.4	Free	Mac OS X	http://www.paranoidfish.org/projects/webkit2png/	*Individual web page in a static state*
WebHTTrack Website Copier	3.33-beta 3	Free	Mac OS X/ UNIX	http://www.httrack.com	*Complete website* Command-line based programme in various UNIX versions (among others for Linux and Mac OS X)
WebReaper	9.8	Free	Windows	http://www.webreaper.net	*Complete website*
Wget (+ wGetGUI 1.05)	1.9	Free	UNIX/ Win-dows/ Mac OS X	http://www.gnu.org/software/wget/	*Complete website* The command-line programme 'wget'. Tested with graphic interface.
WinHTTrack Website Copier	3.33-beta 3	Free	Windows	http://www.httrack.com	*Entire website* Windows-version of WebHTTRack (with graphic user-interface).

For both speed and functionality tests, the computer equipment used was comparable to the ordinary personal computer typically available to the student or researcher. Table 3 shows the most important information on the computer equipment used.

Table 3: Computer equipment used in the test

	Apple Power Mac G4	PC
CPU-type	PowerPC G4	Intel Celeron
CPU-speed.	1 GHz	2,8 GHz
Memory	512 MB	512 MB
Operative system	Mac OS X (ver. 10.3.5)	Windows XP Professional (SP2)

Test of functionality

The functionality test of the individual programme's ability to archive websites was carried out based on a typology of the content-elements of which web pages are constructed. The typology is based on Brügger's theoretical deliberations, for which reason I refer to these for a more thorough explication of the tested elements.[3]

The test results are available on the Centre for Internet Research's website at http://cfi.imv.au.dk/eng/pub/webarc, where overviews, detailed test results and recommendations for use of the individual programmes can be found. All the archiving programmes have been tested with their standard settings, if not otherwise noted in the column headed 'recommended settings'. For each type of content element, there is an evaluation of how well the tested programme is capable of archiving it. The evaluation is both qualitative – as written evaluation – and quantitative – as a grading on a scale of 0-4, as seen in table 4.

Table 4: Grading in the test, based on the following scale, showing proportion of elements archived

Mark	0	1	2	3	4
Evaluation	None	Few	Average	Most	All

[3] The typology can be seen in the detailed reports of the test for functionality (http://cfi.imv.au.dk/eng/pub/webarc/). See also the book *Archiving Websites. General Considerations and Strategies* (Brügger 2005) for the theoretical background for the typology. The book's appendix 1 is a schematic overview of the movable parts of the typology.

Evaluation of the programmes' ability to correctly archive web pages was carried out by validating the archived web pages, compared to the same web pages downloaded to the browser in the normal manner. The validation took place immediately after ended archiving, and the browser into which the archived material was loaded was set to not use cache (cache was set at 0 MB). Furthermore, there was no connection to the Internet at the time of validation (the cable was physically disconnected). These necessary measures were taken because content elements which had not been archived, but which could be found either in the browser's cache or online, could otherwise have influenced the archived web pages. Validation took place by 'paging' between windows with the online version and the archived version of the same web page, in order to evaluate the extent to which the archived web page's content elements and functionality were included in the archiving. See also Brüggers book, where validation of archived web pages is one of the subjects treated (Brügger 2005: 44).

Experience from the functionality test

Execution of the functionality test brought to the fore a number of general pragmatic and technical problems. These will be discussed in this section to the extent they have not already been treated by Brügger (Brügger 2005).

One of the most important things we have learned is connected with archiving using programmes for 'archiving a complete website' (or parts of one). It was discovered that it can be difficult to limit archiving to the web pages desired. The archiving was either too extensive or too limited, without archiving enough content elements to be sufficiently useful for research or term papers. This type of archiving software can be limited by either 1) determining the depth to which the software is to archive (the number of levels of links to be followed in the website's hyperstructure), 2) determining the domains the software is allowed to move in (among other things, it is possible to determine how far the software can be allowed to move outside the website's domain, if at all, while following links from the website), or 3) by using a 'filter' with web addresses that include or exclude respectively, addresses in/from the archiving. Depending on the purpose of the archiving and the complexity of the website, we have learned that this problem is best solved by combining all three types of limitations. This is especially the case when archiving a very complex website such as http://www.dr.dk. It will usually be advisable to begin by analysing the hyperstructure of a website to be archived, in order to exclude the parts of the website that are unequivocally irrelevant for subsequent use in research. Regardless of what software for archiving complete websites is chosen, it will be an advantage to have a certain knowledge of the website's hyperstructure, in order to be able to limit archiving to the most important elements. Furthermore, it is important to decide the extent to which web pages and other elements external to the website are to be included.

Another point to be noted in archiving a complete website or parts of one is that it is often necessary to archive at least one more level of the hyperstructure than what is to be used in research. For instance if three levels of a website are to be archived, it will often be necessary to archive four, since the last underlying level of the archived website will otherwise be incomplete. These limitations are necessary, at least when archiving complex websites, in that archiving will otherwise easily become extremely voluminous, occupying large amounts of storage capacity, and at the same time requiring long periods of time to carry out, thus worsening the problem of time lag (cf. Brügger 2005: 40).

A third point in archiving a complete website or parts of one is that the archiving programmes occasionally 'freeze' while archiving. It has not been possible to completely eliminate this problem in connection with the test, nor has it been possible to ascertain the cause of the problem. However, it should be noted that many archiving programmes allow archiving processes that have been stopped to resume instead of starting over from the beginning. This can, of course, give rise to problems with time lags in the archived material, but can be seen as the least unsatisfactory solution when an archiving process 'freezes'.

A forth discovery is that it is important for the archiving format to be taken into consideration in archiving – especially when it is a case of programmes for archiving a complete website. Certain programmes archive in their own programme-specific formats, which are not necessarily future-oriented. Examples can be seen in the tested programmes Microsoft Internet Explorer 6.0 for Windows in the .mht format, and Microsoft Internet Explorer 5.2.3 for Mac OS X in the .waff format. These formats can only be loaded into the archiving programme used for archiving, so that any further use of the archived website is dependent on a specific operative system and computer programme. Programmes that archive websites in this way have been included in the test because they are part of the package with Windows and Mac OS X, respectively, and we therefore assume that they are widely used for archiving by many researchers and students. The programmes have certain qualities, but it should be considered whether use of the archived website requires storage and use over a longer period of time, as well as whether it needs to be independent of a specific operative system. If the archived website is to be used for purposes such as appendices or documentation, it should not be a prerequisite that the reader is in possession of a specific type of computer equipment or special software in order to consult the document.

Speed test

Finally, the selected programmes were tested with an aim to measuring the speed at which the selected programmes archived. The speed tests were carried out at the same time of day, and were duplicated to increase the reliability of the test. Test number 1 was done between 6-8 a.m. on weekdays, and test number 2 between 10 p.m. and midnight on weekdays. This test was only partly carried out for programmes archiving a single website in a static state, screenshots and screen recordings, since these methods of archiving save snapshots "instantaneously" and because the speed of screen recording is primarily dependent on the person doing the archiving and less on the characteristics of the programme.

Conclusion

The test of eighteen different computer programmes for micro-archiving of websites, divided into four archiving methods, was carried out in the period from July-October 2004, by MA student Bo Hovgaard Thomasen, on the basis of Niels Brüggers book (2004) *Archiving Websites. General Considerations and Strategies.*

As regards software for archiving an 'complete website', it must be concluded that the programmes from the test that provided the most complete archiving were WebHTTrack 3.33-beta-3 and WinHTTrack 3.33-beta-3. The programmes DeepVacuum 1.24, wget 1.9 and WebReaper 9.8 can also be used, but the archiving processes carried out by these programmes have more defects than those carried out by the two programmes mentioned first. There are considerable differences between the programmes, among other things in archiving speeds, but they are all capable of archiving websites so that they usually appear essentially the same as when experienced online. The exception is that content elements requiring an online connection for viewing cannot be archived using this method. Furthermore, it has been shown to be advantageous to use both link-level limitation and domain limitation (internal/external), as well as filtering in order to ensure that archiving is limited to the desired web pages.

A further advantage of the five above-mentioned programmes is that they can be used free of charge, and are continuously being further developed and updated. The remaining programmes tested for archiving a complete website cannot be recommended. This is either because they are not capable of archiving a sufficient number of content elements, so that the archived web page does not appear in an acceptably correct version, or because the programmes use an operative system- and programme-specific format, or finally, because the purchase cost is high compared to the programmes' capabilities; especially considering that the first five programmes can be acquired free of charge.

All the programmes tested for archiving 'Individual web page in a static state', 'screenshot' and 'screen recording' can be used to archive. Some of the programmes allow for more flexibility than others, for instance with regard to editing or choosing the archiving format, but this is usually reflected in the purchase cost. However, SnagIt 7.1.2 deserves special mention, since the programme is capable of archiving using all three above-mentioned methods, while being extremely easy to use as well as fast. The programme is, however, one of the most expensive in the test. One programme, Web2Pic 1.1 ('Individual web page in static state') is not capable of archiving all types of web pages.

References

Brügger, Niels (2005). *Archiving Websites. General Considerations and Strategies* the Centre for Internet Research, Århus. Available online at: http://cfi.imv.au.dk/pub/boeger/bruegger_archiving.pdf