Archiving Websites

General Considerations and Strategies

Niels Brügger

Center for Internetforskning
The Centre for Internet Research

Published by The Centre for Internet Research, Århus, Denmark,
January 2005.

Niels Brügger: *Archiving Websites. General Considerations and Strategies*
1st edition 2005

Translated by Stacey Cozart and Patricia Lunddahl.
Printed at Werks Offset A/S.
Cover design: Thomas Andreasen.
ISBN: 87-990507-0-6

This book was published with support from the Danish research project MODINET (Media and Democracy in the Network Society) and the Faculty of Humanities, University of Aarhus.

The Centre for Internet Research
Institute of Information and Media Studies
Helsingforsgade 14
DK-8200 Aarhus N
cfi_editors@imv.au.dk
Tel.: + 45 8942 9202
http://cfi.imv.au.dk

Table of Contents

Appendix 2

Step-by-step guide to archiving a website

Since an archived website to a certain degree is only shaped in the archiving, it should be accompanied by a document containing methodical considerations of why and how the website has been archived. The following step-by-step guide is meant as an aid to the outline of such a document. In addition, it will naturally also act as a practical aid in connection with the actual archiving (and, of course, the following is to be seen in the context of the previous pages' general

deliberations and strategies, which it condenses in an itemised, tabular form. The guide is divided into two main parts: 1) prior to archiving, 2) the archiving process. An electronic version of this step-by-step guide can be found at http://cfi.imv.au.dk/eng/pub/webarc.

Prior to archiving

The analytic purpose

It is important if possible to clarify the analytic purpose the archived material is intended to serve. If the analytic purpose is unclear or unknown, one must be aware that it may be impossible to use the archived material in a later analysis.

Also, it should be considered whether the analytic purpose can be served without archiving the website (for instance, by observing and documenting web-site activity using quantitative or qualitative methods).

Existing archives

One should attempt to ascertain whether the desired website is archived in any of the existing national or international Internet archives from the desired period, in a form and of a standard wholly or partly suited to the analytic purpose. If desired, a list of larger international initiatives related to Internet archiving can be seen at http://www.nla.gov.au/padi/topics/92.html.

Supplying by the producer

Investigate the possibility of obtaining a copy of the website from its producer. Clarify as quickly as possible any technical, organisational, financial, temporal and copyright-related questions with regard to delivery.

Ethical and regulatory questions

Consideration should be given to any ethical and regulatory questions regarding the website to be archived, the actual process of archiving, how the archived material is made accessible, and so on.

Types of archiving software

Consideration must be given to the type(s) of archiving software to be used in order to fulfil the analytic purpose: a) will we use software that can archive entire websites or single web pages in a static form, which can make a screenshot or screen recording? b) will we use a combined form with regard to documentation, exemplification or contextualisation? C) will we use an 'imprecise' analytic strategy because the analytic purpose is unclear or unknown? The following tables can help answer these questions.

Basic types of archiving software[1]

Archiving software for entire websites				
can archive	cannot archive	Important considerations in regard to:		
		space	time	montage
the structure of the website its elements the ability to move freely between them movement of elements where archiving does not require continuous online presence	movement in elements where archiving requires continuous online presence	size and complexity file size typically in megabytes	the dynamic of updating theoretically no need for personal presence (but often required in practice)	archival elements often imprecisely delimited with regard to space and time

Archiving software for static archiving of individual web pages, or for screen shots				
can archive	cannot archive	Important considerations in regard to:		
		space	time	montage
individual webpages	any form of action, between or in elements	file size typically in kilobytes	personal presence required during entire archiving process	archival elements precisely delimited with respect to both time and space

1. If only sound is to be archived, it is advisable to use software developed exclusively for recording sound (for instance, WireTap (for Mac: http://www.ambrosiasw.com/utilities/freebies).

Screen-recording software				
(with movement on the website)				
can archive	cannot archive	important considerations in regard to:		
		space	time	montage
website structure and elements individual elements where activation of movements requires continuous online presence	the ability to freely move in website structure	size and complexity file size typically in megabytes; file size increases rapidly with image quality	personal presence required during entire archiving process	archival elements precisely delimited with regard to time and space
(without movement on the website, for instance movement in one element alone)				
can archive	cannot archive	Important considerations in regard to:		
		space	time	montage
website structure and elements individual elements where activation of movements requires continuous online presence	the ability to freely move in website structure	file size typically in megabytes; file size increases rapidly along with image quality	personal presence theoretically not required during archiving process, but as films will often be short since file size increases very rapidly, a person will usually be present during the archiving process	archival elements precisely delimited with regard to time and space

Combined forms[1]

	Use	Consists of	Space and time
documentation	to ensure that all elements of expression are actually archived, and that they are archived as they appeared and were placed to archive changes to a website noted during archiving and considered material for one or another reason	archiving software for archiving of entire websites + either static images or screen recording	small website: may choose to document in entirety larger website: must choose specific areas — for instance, those central to a subsequent analysis or web pages that are especially important in general (navigational 'crossroads', pages with high updating frequency, etc.)
exemplification	to be able to show examples of specific movable elements or types of elements requiring continuous online presence, without, however, making these the object of a thorough analysis	archiving software for entire websites + screen recording	website size less important important to know where in the structure the movable element to be exemplified is to be found
contextualisation	to archive the website as a seamless textual system of expression — i.e. as *both* structure, elements, the ability to move between elements *and* the possibility of movement in all elements, including those where activation requires continuous online presence; is carried out with the aim of allowing for a subsequent thorough analysis	archiving software for entire websites + screen recording	website size less important important to know where in the structure the movable elements to be archived is to be found

1. In all three combined forms treated, it is recommended that the two archiving processes be carried out simultaneously by beginning the archiving of the structure and archiving individual images or film while the structure is being archived (this, however, is dependent on sufficient memory and storage capacity).

   In all three combined forms one must be aware of the fact that montage becomes more complicated, in that we now have a *third montage* where two different types of archiving of 'the same thing' are subsequently to be combined.

Four 'imprecise' strategies

Website size	Little time available	More than sufficient time available
small	use archiving software for archiving of entire websites or static images If possible: • draw up site diagram • document central web pages • search for elements requiring user intervention and online presence There will probably not be time to check the quality of the archived material	• use archiving software for archiving of entire websites • draw up site diagram • document central web pages • consider combining the two remaining combined forms • search for elements requiring user intervention and online presence There should be time to check the quality of the archived material.
large	use archiving software for archiving of entire websites	use archiving software for archiving of entire websites If possible: • draw up site diagram • document central web pages • consider combining the two remaining combined forms • search for elements requiring user intervention and online presence There will probably not be time to check the quality of the archived material.

Specific means of archiving

Parallel with deliberations over the type of archiving software to be used, it must be clarified whether the necessary means are available to carry out archiving as planned. In this connection, one should consider the following questions, which are mutually dependent:

- What specific software of each type is to be used? (the following play a role here: archiving quality, speed, price, user-friendliness, documentation)
- Are several different archiving software programmes of the same type to be used?
- What hardware is available (platform, processor speed, working memory, storage capacity (both during and after archiving))?
- The person doing the archiving (Can the archiving software be easily used? Is there time enough, and is the person able to become familiar with the use of the software in question? Is archiving software affordable?)

For a discussion of several of these questions, see the test results available at:

http://cfi.imv.au.dk/eng/pub/webarc.

Site diagram

Regardless of the type of archiving software used (alone or combined), a site diagram should be drawn up describing the structure of the entire website or the part of it to be archived (this is not, however, as necessary in screen recording, when movement is in one and the same element). The larger the website, the more essential it is to draw up a site diagram.

There are two joint goals for the site diagram: partly it is to provide an overview of the structure of the website, partly it is to act as an archiving log, which can later be used as documentation in general as well as documentation of any sources of error. Thus it can be used before and during archiving for notes regarding the following points and any changes to be made in them:

- the choice of archival elements and direction, including whether (and if relevant, where) several archival elements are to be archived simultaneously
- areas of special difficulty, such as navigational 'crossroads', areas with high updating frequencies, or semantic relationships between several areas with rapid rates of change
- where passwords are required and whether they have been procured (remember that in screen recording, logging-in will be shown on the screen film)
- where the type(s) of archiving software are used on the website
- what has been archived, what has been validated, both in relation to the website existing on the Internet, and in relation to the website in the archive

The site diagram can be drawn up based on the menu structure of the website, a site map, etc. It should be created as close to the time of archiving as possible.

Anticipation of subsequent treatment (storage, searching, viewing)

Before beginning archiving, the following questions regarding subsequent use of the archived material should be considered: How is material to be stored, searched and viewed?

Storage

- Is the material to be compressed? (takes up less space, but makes access more difficult).
- How can files be stored, and not least named, in a clear, well-structured manner, so they are not mixed and so the chosen strategy is mirrored in some fashion?

It is recommended that a net archive be organised as soon as possible after archiving, in that the number of files can quickly become impossible to manage.

Searches

- Should it be possible to perform a direct full text search of the archived material? (this is usually difficult, but with the type of archiving software that saves in single files like those on the website being archived (such as HTTrack), with Macintosh it is possible to perform a full text search by searching for file content using 'Find' in 'Finder').
- Is some form of overview with a list of files to be created, and how is it to be organised?

Viewing

- Is the archived material to be accessible off- or online? by only one or by a number of persons?
- Does viewing the files require special software?
- Can the archived material be viewed on all platforms?
- Should it be possible to view the material in 10 years time? (if so, a widespread file format should be chosen)

The archiving process

If possible, one or more test runs should be done before beginning the actual archiving process.

Archiving software for entire websites

- The entire website or the desired area(s) of the site is archived
- the archived material is validated (compared to the website in existence on the Internet)
- if the quality is not acceptable, archival elements and direction are determined (however, it can be an advantage to save the archived copy of the entire website, as it (in spite of errors) may be used to document the structure of the website)

The following steps can be taken sequentially or parallel, depending on whether we are archiving one archival element or several at a time.

- archival element 1 is archived
- archival element 1 is validated (compared to the website in existence on the Internet), and any necessary adjustments to archival elements and direction are made[1]
- archival element 2 is archived

1. In connection with validation compared to the existing website on the Internet, one should be especially aware of two possible sources of error. Firstly, the archived website, often without warning, can obtain material from the active Internet if the connection is open — which it is, since the archived material is being compared to Internet material. Secondly, even if the connection to the Internet is closed, it will be possible for the archived website to obtain material not actually in the archived material, but temporarily stored on the computer in the so-called cache (an area of the hard disk or memory where recently visited pages are stored). These two problems can best be solved as follows: open a browser showing the active website on the Internet and open another browser operating offline, in which you can view the archived website; this browser should be set to empty its cache and/or to a disk cache value of 0 MB, and the memory cache is set to never compare a website to cache. Otherwise, validation is best done by opening two windows of the same size, laying them one on top of the other, and then paging down, shifting, paging down, etc.

- archival element 2 is validated (compared to the website in existence on the Internet *and* to the archived material), and any necessary adjustments to archival elements and direction are made
- etc.

Archiving software for archiving of individual webpages in a static form or for screenshots

- The archival element is determined (depending on whether one needs to archive entire web pages or parts of web pages/several open windows) and suitable archiving software is chosen
- an archiving direction is determined
- archival element 1 is archived (if necessary validating, archiving again, and correcting the direction when using software for archiving single pages)
- archival element 2 is archived
- etc.

Screen-recording software (with movement on the website)

- The archival element and direction are determined
- archival element 1 is archived (assessing whether the archival element and direction are still appropriate)
- archival element 2 is archived (assessing whether the archival element and direction are still appropriate)
- etc.

Screen-recording software (without movement on the website)

- The context may be archived (may be done separately)
- the element in question is archived (remember to close down any screen savers before beginning archiving)

It should be noted that saving after recording is time-consuming in both types of screen recording.