Niels Brügger

Archiving Websites

General Considerations and Strategies

Archiving Websites

By the same author

Jean-François Lyotard: Sprog, teknik, filosofi, Aarhus University Press, Århus 1989

Filosofiske forskydninger: En bog om Jean-François Lyotard (ed. with F. Frandsen), Akademisk Forlag, Copenhagen 1989

Lyotard og striden: Læsninger af Le différend (ed.), Slagmark, Århus 1990

Lyotard, les déplacements philosophiques (ed. with F. Frandsen, D. Pirotte), De Boeck-Wesmael, Bruxelles 1993

Paul Virilio: Krigen, byen og det politiske (ed. with H.N. Petersen), forlaget politisk revy, Copenhagen 1994

Foucaults masker (ed. with K.O. Eliasen, J.E. Kristensen), Modtryk/Spartacus, Århus/Oslo 1995

Virilio: Essays om dromologi, Introite!, Copenhagen 2001

Media History: Theories, Methods, Analysis (ed. with S. Kolstrup), Aarhus University Press, Århus 2002

The Internet and Society? Questioning Answers and Answering Questions (ed. with H. Bødker), The Centre for Internet Research, Papers from The Centre for Internet Research, 05, Århus 2002

Strukturalisme (with O. Vigsø), Samfundslitteratur/Roskilde Universitetsforlag, Copenhagen 2002 (Swedish edition: *Strukturalism*, Studentlitteratur, Lund 2004)

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

Preface 7

1. Micro and macro archiving 9

2. Document, monument and imprint 15

Document 16

Monument 17

Imprint 18

3. The dynamics of the Internet 21

Sender 21

The dynamic of updating 22

The dynamic of proliferation 24

Text 25

The dynamic of movement 25

The dynamic of complexity 26

Recipient 26

The dynamic of equipment 27

The dynamic of actions 27

4. Archiving the dynamics of the Internet? 29

Document, monument or imprint? 29

A document of the Internet 30

The need for considerations of method 31

5. Test of archiving software **33**

A completely archived website? 33

Types of archiving software 34

Prerequisites and results 34

6. Elements of an archiving strategy **37**

Building blocks and variables 38

Building blocks: types of archiving software 38

Variables: space, time, montage 39

Building blocks and variables 41

Combined forms and purposelessness 53

Combined forms 53

Purposelessness 58

7. Representation and subjective involvement **61**

Bibliography 63

Appendix 1: Typology of movement in elements

Appendix 2: Step-by-step guide to archiving a website

Preface

The following text was written within the framework of the Danish research project MODINET (Media and Democracy in the Network Society, 2002-05, www.modinet.dk), and constitutes (necessary) groundwork for a future study of www.dr.dk, the website of the Danish public service broadcaster. The practical realisation of the software test that comprises part of the subject matter of the book was made possible by financial support from MODINET, and it was carried out by graduate student Bo Hovgaard Thomasen.

I am grateful to a number of people who have contributed to this book with relevant comments and criticism: Bo Hovgaard Thomasen, Rikke Agersnap, Niels Ole Finnemann, and the members of MODINET's team 3. The author nevertheless bears full responsibility for the text that follows. Finally, I thank the University of Aarhus Research Foundation for placing Møllehuset on the Sandbjerg Estate at my disposal during the writing phase.

In connection with the publication of this book a website has been established where it is possible to find the conclusions of the test, the detailed test results, recommendations for using the individual programmes, a detailed description of the test, and links to resources on net archiving. The website address is http://cfi.imv.au.dk/eng/pub/webarc.

Aarhus & Sandbjerg, October-November 2004

Chapter 1

Micro and macro archiving

The reason why Internet research even concerns itself with archiving the Internet is that at some point research that has the Internet as its concrete object of study needs to stabilise and maintain this object in order to preserve it, either for immediate use in an analysis and/or for later documentation and thereby as a basis for criticising and discussing the analysis performed. The question of Internet archiving is thus situated at an intersection between research and archiving in the broadest sense.

National and international archiving institutions such as national libraries or (semi-) private organisations will to a great extent best be able to attend to archiving the Internet.[1]

---

1. In Denmark, a national Internet archive is currently being established, and similar national Internet archives exist in a number of other countries. In the US, there is archive.org, which is a non-governmental Internet archive. For a combined list of the major Internet archiving initiatives launched by national libraries and the like, see PADI, Preserving Access to Digital Information, http://www.nla.gov.au/padi/topics/92.html. Finally, in July 2003, the 'International Internet Preservation Consortium' (IIPC) was set up: an international agency whose goal is "a) collaborative working, within each country's legislative framework, to identify, develop and facilitate implementation of solutions for selecting, collecting, preserving and providing access to internet content; b) facilitating international coverage of internet content archive collections within national legal frameworks and in accordance with individual national collection development policies; c) international advocacy for initiatives that encourage the collection, preservation and access to internet content" (IIPC 2004). Goals that can be achieved by "a) provid[ing] a forum for sharing knowledge about internet content archiving both within the Consortium and beyond; b) develop[ing] and recommend[ing] standards; c) develop[ing] interoperable tools and techniques to acquire, archive and provide access to web sites; d) rais[ing] awareness of internet preservation issues and initiatives through conferences, workshops, training events, publications, etc." (ibid.).

However, the individual researcher, student or other person with an interest in the Internet who wants to use an archived version of a particular website will not always be able to count on finding it in an archive in the desired form and version and as quickly as desired. If archiving is done selectively (selected websites, Internet activity in connection with particular events and so forth), the website may not be archived often enough (or at all), or it may not be archived from a specific day if archiving is carried out according to the principles of cross-sectional harvesting (e.g. all of .dk, which takes several months), or perhaps it isn't accessible as quickly as one needs it (e.g. if the archived material needs to be treated first).[1] And unless one is a researcher and has the time to continuously seek the necessary permission, it may not be possible to access the archived material at all.

This book distinguishes between two general approaches to Internet archiving: *micro* and *macro archiving*.

Micro archiving means archiving carried out

- on a small scale, both as regards space (a limited number of websites) and time (a limited, isolated period)
- by individuals who do not have at their disposal considerable computer power and storage capacity and whose technical knowledge of archiving or of the subsequent treatment is either lacking or on an amateur level
- on the basis of an immediate, here-and-now need to preserve an object of study

1. Several of these circumstances are described in the reports prepared in connection with the Internet archiving project "netarkivet.dk"; see Brügger et al. 2003, and the interim reports from the project, which can be downloaded from netarchive.dk.

In contrast, macro archiving is carried out

- on a large scale (of a large number of websites and in principle infinitely)
- by institutions (public/private, national/international) that have considerable computer power, storage capacity, and professional technical expertise at their disposal
- in order to archive (part of) the (inter)national cultural heritage.[1]

The topic of this book is the *micro archiving of websites*. It is therefore primarily addressed to researchers, students or others without specific technical knowledge who occasionally want to preserve a website for purposes of study by means of their standard PC or Mac. Incidentally, the purpose of micro archiving can be something other than research — such as being able to document for a purpose such as legal proceedings and the like what was actually on a particular website at a given point in time, especially when the website has since been changed or removed. This might be in connection with everything from threats of international terrorism to illegal web publications and all kinds of civic disputes.

This book focuses therefore on the website and not on the Internet in all its other forms (usenet, etc.). The term 'website' is defined as a coherent unit based on its content and editorial framework rather than as regards its technical characteristics.[2] Several of the general considerations apply to the archiving of the Internet in general.

Likewise, the book does not focus on archiving with regard to long-term preservation, with all the resulting problems with respect to deterioration of

---

1. The importance of preserving the part of cultural heritage comprised by the Internet is among other things discussed in Finnemann 2001a, Finnemann 2001b and Merzeau 2003: 161-163.

2. Cf. Niels Ole Finnemann's discussion in Finnemann 2005 (pp. 171-174), which results in the following definition: "A website can be defined as a site, 1) i.e., *a delimited set of addresses* on the Internet, which is delimited insofar as 2) they are *subject to an overall editorial control* of their *content*, which is 3) *freely accessible to the general public* either through payment or free of charge and with or without user indication and a password" (p. 173).

storage media, disappearance of techniques, programmes, operational systems, file types, and so on.[1]

Finally, the book focuses on archiving, but not on the supplying of the website by the producer, with the specific issues related to this. At first glance, it appears obvious to approach the producer of the website one wants to study in order to obtain a copy of it. This can be practicable, and in certain cases should be tried, but will usually lead to various kinds of intertwined problems, including technical ones (it will very likely be necessary to set up a technical environment (hard/software) in order for the website to function (cf. Brügger et al. 2003: 22), and it can also be difficult to 'collect' the different parts of the website), organisational ones (Who is authorised to answer a request for supplying a website? Who should be responsible for supplying it?), copyright issues (Should everything be supplied, or should the material be sorted through? And if so, who should do it?), economic issues (Who should pay for the preparation and realisation of its delivery?), and temporal issues (Settling all the above problems takes time). The technical as well as the organisational, copyright, economic and temporal problems tend to become greater the larger and more complicated the website in question is, and in connection with delivery there is a de facto risk of moving further and further away from micro archiving (at a minimum, a certain degree of technical insight is often necessary). In practice, the delivery will thus by and large have to be left to bigger (inter)national archiving institutions, typically based on prior agreements regarding technical specifications, supply dates, costs and so forth.

Finally, it should be mentioned that a number of ethical and regulatory questions will often have to be considered in connection with Internet archiving. These will not be further discussed in this context; it should merely be noted that with regard to general ethical considerations of Internet research, inspiration can be found in a publication issued by the Association of Internet Researchers, "Ethical decision-making and Internet research: Recommendations

1. For additional information on long-term preservation, see Brügger et al. 2003: 22, 29-30, 49.

from the aoir ethics working committee" (Ess 2002). With regard to conditions more specific to archiving, one question to be considered is whether the archiving should comply with a website's specifically expressed desire not to be archived (expressed in the file robots.txt, which certain archiving software can be set to accept or not). And finally it should be emphasized that according to many countries' legislation, archived Internet material is intended for own use only, which is why it may not be published on the Internet. An archive therefore ought primarily to be established on one's own machine, and in any case it is advisable to become familiar with existing copyright laws in connection with Internet archiving.

Chapter 2

Document, monument and imprint

Based on the fact that at some point Internet research needs to stabilise and maintain its object so that it can somehow be preserved, the question naturally arises of whether the Internet *is* always already preserved — isn't it always 'out there'? It turns out that the material on the Internet changes — or is removed — rapidly: Forty percent of the material on the Internet disappears within a year, while another forty percent has been changed, which is why today we can only expect to find twenty percent of the material that was on the Internet one year ago. The Internet is in other words a very *dynamic* object.[1]

The Internet is obviously not the first dynamic object of research, but it is

1. Some of the material that does not change is presumably not meant to. These may be actual archives (either individual websites or parts of a website), usenet groups and weblogs, or websites that have been preserved on the Internet as they were at a particular point in time for historical reasons. Examples of the last category include the Mosaic browser's original documents from 1993 and onward, which have been preserved so that one can learn that "The basic means of navigation is clicking", and that one should "Click on the coloured text to follow a link" (see http://archive.ncsa.uiuc.edu/SDG/Software/Mosaic/), and furthermore the original versions of NCSA 'What's New' (1993-1996), which are web pages with lists of links to new websites (see http://archive.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/whats-new.html). This kind of archive on the Internet, which may be regarded as a kind of website archive, is made available by the author, which however means that this person will be able to change, move or remove them, sell them or the like, and possibly at some point the technical infrastructure will not even allow them to function (an example of this is the PARC MAP Viewer from Palo Alto Research Center, http://pubweb.parc.xerox.com:80/map). This way of archiving the Internet, by means of which the author archives in an 'open' archive, therefore distinguishes itself from the Internet archiving carried out by larger archiving institutions whose main task is to archive the Internet as well as to ensure that the archived material does not change; the former can be called a kind of 'mezzo archiving'.

dynamic in a number of distinct ways — beyond the fact that its content quickly changes/is removed — that need to be taken into consideration in connection with its archiving.

Before elaborating on these particular dynamics of the Internet, I shall outline the general archival frame of understanding in which they are situated — a frame based on how one otherwise attempts to make a dynamic object stable and thus possible to preserve for purposes of research. How this stabilisation and preservation takes place depends, of course, on the type of concrete analytical object and the sense in which it can be considered dynamic. It should be emphasized that the dynamics outlined in this chapter are exclusively motivated by the general aim of this book: to discuss archiving. Therefore, it will of course be possible to indicate other dynamics than those emphasized below.[1]

I shall distinguish between three general categories for stabilising and preserving dynamic objects: *document*, *monument* and *imprint*.

Document

One type of concrete object of analysis is dynamic by being highly complex and changeable. This could for instance be nature or human beings and their social behaviour, actions, and thoughts in the broadest sense.

The dynamic of the concrete object of analysis can be stabilised and preserved for research by its being *transformed* with the aid of some sort of symbolic system, reproduced in some sort of medium that is different from the object itself. Take, for instance, the natural scientist, social scientist, psychologist, or ethnographer who attempts to capture and secure nature, society, the psyche, culture and so on by observing, monitoring, posing questions, keeping records and taking notes, interviewing, photographing and filming the concrete objects of analysis.

The result is a collection of *documents about* the object, which can be

1. I shall furthermore only discuss the preservation of objects found in the present rather than preservation in a historico-archaeological sense — i.e. archiving objects from the past.

preserved, while the object itself is not preserved. Depending on the symbolic system and type of media, these documents are to varying degrees 'iconographically similar' to the object — from written characters, numbers, various systems of notation and graphs to sound and pictures, each of which can occur on paper, film strips, sound/videotapes and in various digital formats.

The transition from dynamic to stable here requires the involvement of a subjective consciousness, partly in selection, partly in various ways in the transformative process itself.

Monument

Another kind of concrete object of analysis is dynamic solely because it is used and circulates among people. This could for example be all the objects with which we surround ourselves, and that exist in society in a material form that can be directly preserved, from everyday domestic utensils to types of media such as the printed media, photographs and film.

Here, the dynamic can be made stable and preserved simply by taking the concrete object out of circulation and *preserving* it in toto, unchanged. Linguistic and cultural researchers, art historians, literary and media researchers and others either preserve their objects of study themselves or go to see them in places where they are out of circulation (libraries, museums, archives and so on).

The result of this procedure is a collection of *monuments* that are the objects.

In this case, the transition from dynamic to stable only requires the involvement of a subjective consciousness in selecting the exemplary elements, while the actual preservation process is merely the 'taking out of circulation'; regardless of who preserves the Viking sword at the National Museum or who stacks the old newspapers at the library, they look the same — what is preserved is identical to what circulated.

Imprint

A third type of concrete object of analysis is also dynamic, in that it circulates in society, but circulates in a material form that cannot be directly preserved. These are things like the transmitted, 'immaterial' objects with which we surround ourselves, from radio and TV broadcasts to cellular telephone services and the like.

Here, the dynamic element can be stabilised and preserved simply by taking the concrete object out of circulation and *recreating* it by transferring it from the type of medium in which it appears to another. Broadcast audio-visual material is for instance recorded and preserved on tape, DVD, hard disk and the like.

The result is a collection of *imprints* bearing witness to the object.

Again, in this case the transition from dynamic to stable only requires the involvement of a subjective consciousness in selecting the exemplary elements, while the actual process of preservation is more mechanical: the TV programme looks the same regardless of who puts in the tape and presses the record button — what is preserved is identical to what circulated, but it is preserved in another kind of medium.

Three different ways of stabilising the dynamic features of objects of analysis are thus under consideration. The first creates *documents* about its objects, the second stores *monuments,* which are the objects, while the third stores *imprints* of the objects. If we compare the three, the following differences and similarities become evident:

- In the first case, the dynamic of the object is characterised by a high degree of complexity and changeability, while for the other two it is found solely in the object's circulation in a broad sense. Depending on the kind of object, the ways in which it circulates and is used can, of course change it, but as far as the kinds of media are concerned, they seldom change once they have been put into circulation — the newspaper, book, photograph, film, radio and TV do not change once they have been 'published'.

- At first, the stabilisation of the object in the shape of a document requires a *symbolisation/mediation* of the object, which the monument and the imprint do not (the starting point of the monument can, for instance, already be a medium, while the object of the imprint is itself a medium that is merely 'remediated' in a weak sense of the word).
- The document and the imprint, unlike the monument, involve an element of *representation* of the object, insofar as it is a matter of 'creating' in some way or another through the act of preservation; however, as regards the imprint this is a very 'weak' kind of representation.
- The document involves a high degree of *subjective involvement* in the very process of stabilisation, which is not true to the same extent for the monument and the imprint.

Chapter 3

The dynamics of the Internet

The Internet is a dynamic medium in that its contents change or are removed quickly. But the Internet is also dynamic in a number of other ways that on a general level are characterised by their complexity, their high speed, and their unpredictability. To be able to treat these general conditions systematically, and to further elaborate what more precisely is implied by the changing contents of the Internet, one can focus on how they play out on the individual communicative elements: the *sender*, *text* and *recipient*. The dynamic features apparent in these three are all made possible by the fundamental mode of existence of the medium — the Internet — itself: it is composed of computers in networks.[1]

Sender

It is possible to distinguish between two dynamic aspects of the sender: the *dynamic of updating* and the *dynamic of proliferation*.[2]

---

1. The dynamic features of the Internet outlined in the following are not an exhaustive list, but rather some of those that play the most important role when the topic is the archiving of the Internet, and in other contexts it would probably be possible to point out others. It should also be noted that the dynamic features are temporally bound insofar as they are what exist today; this book would have been different ten years ago, and it would most probably be different in five-ten years as the Internet changes. Finally, it should be mentioned that the elements of the sender, text and recipient can in certain cases be difficult to keep separate, in that the medium 'the Internet' in many ways interweaves them. The analytical lines may therefore be drawn more sharply than is necessary.
2. The dynamic features of the sender are discussed in Brügger 2001. The text about the sender in the present book is in parts identical to Brügger 2001: 45-51 (with minor editorial changes)

The dynamic of updating

Updating — the fact that a sender renews the contents of a medium or makes new versions — concerns space and time. To be able to archive the updates we must know where and when they occur. In this respect, the Internet constitutes a break with the high predictability and the one-dimensional temporal logic characterising past media. Firstly, the updated version is not materially distinct from, for instance, that of the day before, rather it is situated 'inside' the same medium. In other words, we never know precisely *where* the Internet is updated. Secondly, updating does not follow a one-dimensional temporal logic (a publishing rhythm, a programme, a time schedule or the like), but rather a multi-dimensional temporal logic. We never know in advance precisely *when* the Internet is updated.

To be able to ensure that the object we have in our archive is identical to the object as it really was in the past, we must be able to control the updates of the sender, both as regards space and time. If this is not possible, we risk studying an object that in some way or another is not identical to the object as it really was.

This somewhat cryptic wording can be illustrated by the following example, which stems from a very specific Internet genre, news media on the Internet; but most websites have the same characteristic: they are updated often and irregularly, and one never knows where. Even though the example is based on slow and not-very-sophisticated software, it illustrates a general problem of a more fundamental than technical nature. During the Olympics in Sydney in 2000, I wanted to save the website of the Danish newspaper *JyllandsPosten*. I began at the first level, the front page, on which I could read that the Danish badminton player Camilla Martin would play in the finals a half hour later. My computer took about an hour to save this first level, after which time I wanted to download the second level, "Olympics 2000". But on the front page of this section, I could already read the result of the badminton finals (she lost).

---

and parts contain new material.

The website was — as a whole — not the same as when I had started; it had changed in the time it took to archive it, and I could now read the result on the front page, where the match was previously only announced. The example illustrates that the time it took to update the page did not coincide with the time it took to 'publish' the website as a whole — on the Internet the updating has several temporal logics. We cannot be sure that the website we began downloading is identical to the website as it appears only one minute later — akin to the beginning of a TV programme changing when we are halfway through it, which would force us to start over and over again with our archiving. And we never know *where* the changes will occur — or *when*. The result is clear: what is archived is not always identical to everything that has been published.

This has two consequences. The first and rather obvious consequence is that we cannot be sure that we have everything in our archive. We will always have lost something in the asynchronous relationship between updating and archiving. The second consequence is less obvious, but no less serious. Not only do we lose something that was there, we are also in danger of getting something that in a way was *never* there — something that is different from what was really there. My archived version of the newspaper's website can be *a combination* of elements from two (or more) versions that were there at different times — but they were never there at the same time as they might now be in my archive. We thus face the following paradox: on the one hand, the archive is not exactly as the website *really* was in the past (we have lost something), but on the other, the archive may be exactly as the Internet *never* was in the past (we get something different).

The example illustrates a general problem, and in this sense it is — fundamentally — unimportant whether we speed up the archiving process; the asynchronous relationship between updating and archiving remains — like the hare, we will never be able to catch up with the tortoise.

In continuation of this condition, serious problems appear on the horizon as regards issues such as source reliability and references, and source criticism appears to have acquired new, perhaps unfeasible, responsibilities.

Medieval philologists can, for instance, be certain that the sources they study — hand-written manuscripts — have in fact been in circulation and are now in their keeping in the same form. Their task is thus 'merely' to find out which versions are the 'originals' and which are copies. Even though the job can be difficult in practice, it is theoretically possible to determine the relative original/copy relationship between two or more sources. But with an Internet archive we cannot, as mentioned above, be entirely certain that the source we have in the archive actually ever existed on the Internet in the same form. If two or more archives contain the same website — archived at the 'same' time — it is possible, indeed even probable, that they are not identical (cf. above); however, the relationship between the versions can hardly be considered an original/copy relationship, but is rather either a copy/copy or original/original relationship — and theoretically we have no way of clearly determining which it is. It may be a matter of different copies of 'the same', but given that we cannot expect to ever find an original of which they are all copies — what was actually on the Internet at a given point in time — it is difficult for us to determine the reliability of the copies. Or perhaps we are left with a series of different originals — i.e. versions of something that has never been on the Internet but that is in our archives.

Our archive may therefore consist of a confused jumble of copies and originals, and our usual source-critical procedure is not sufficient to draw a clear line between the two.

The dynamic of proliferation

Taking a closer look at the second dynamic aspect of the sender — the dynamic of proliferation — it may be observed that the emergence of senders is characterised by a different dynamic in the case of the Internet than in that of other media. In other media senders come and go, but the Internet is different as concerns their number, mode of existence and speed.

Firstly, anyone who places something on the Internet is a sender, which is why the number of senders is enormous. Secondly, the appearance and disappearance of senders is unpredictable: we never know how or when new senders

will arrive on the Internet — and when they will disappear again or move to other parts of the Internet. Furthermore, they constitute a highly diversified mass, as opposed to senders of familiar kinds of media, among which the variations are fewer, such as publishers and radio/TV stations. Third, the speed of emergence/disappearance/removal is very high and can for instance be linked to certain events that seem to accelerate the appearance of senders: 9/11, elections, natural disasters, wars and so on — or event-like radio/TV programmes like 'Big Brother' and 'Robinson'. The very appearance and spread of senders thus constitutes — for these three reasons — a dynamic element of the Internet.[1]

Text

If we look at the textual elements of expression characterising websites today, some dynamic features are also manifest here: a *dynamic of movement* and a *dynamic of complexity*. The term 'textual elements of expression' is used in the broad sense of written as well as (live) images and sound.

The dynamic of movement

The dynamic of movement refers to two different conditions: in a broad sense, the movements that are evident *between* and *in* the individual elements in the textual structure of the website (structure in the sense of an organisational principle for the individual page and for the multitude of individual pages).

The movement *between* the individual elements is what happens when one is moved from A to B. It can be to pursue a link to a new place on the website or to a place that leads the user partially out of the website — for instance, a link that downloads a file or opens a new window (and then perhaps a plugin/new application), or it can be to be automatically transferred to another website/place on a website. The former requires user intervention, while the latter

1. A decisive element in the archiving of Internet activity in connection with events such as political events, wars, sporting events and so forth is thus the continuous monitoring of newly arrived websites, which is both time-consuming and labour-intensive work.

is an automated movement.

The movements *in* the individual elements are the fact that A or B as a delimited element moves or can move (for a schematic overview of movements in elements, see appendix 1). This might be movement in writing (e.g. scrolling texts, chat), movement in images (animations, graphics, live images) or 'movements' in sound (playing back sound);

- The movements of the elements can either be automated (i.e. they proceed on their own — e.g. animated advertisements) or based on user intervention, which means that something has to be activated. This activation can range from very general with very few alternatives (pressing a start button), to more specific with several alternatives (choosing a programme, camera angle, input in formulas, search engines (own words/pull-down menus), clickable maps …) and, finally, highly specific with an almost endless number of possibilities (e.g. participation in games or chat).
- The possibility of activating the movements of the elements can either be inherent to the element (e.g. clickable maps, sound, image or game files that can be downloaded…) or require a continuous online connection (e.g., participating in chat, streamed sound/video, certain games …).

The dynamic of complexity

The dynamic of complexity refers to the dynamic that arises due to the *synchronous* (and often directly integrated) presence of dynamic text elements; in other words, the dynamic that characterises the actual interaction between, respectively, movements between and movements in the individual elements such as the synchronous presence and seamless integration of various formats of expression (writing, still/live images, sound), several windows/plugins/applications, possibilities for pursuing links, and so on.

Recipient

The reception of a website is characterised by the *dynamic of equipment* and the *dynamic of actions,* respectively. These refer to the fact that the reception

of the same website may differ depending on the recipient's equipment and settings as well as his current and previous actions.

The dynamic of equipment

Firstly, a particular website is sent/shown differently according to the kind of machine that receives it (PC/Mac/cell phone…), which browser is used, how it is set (font size, link colours, opening of plugins…) and which plugin/application versions are employed.

The dynamic of actions

Secondly, the website appears differently according to what the recipient is *currently* doing, from moving/scaling windows (which are not necessarily 'seen' as defined by the sender) to completing forms and so forth, and to what he has *previously,* knowingly or unknowingly, done — for instance, the website can be personalised to various degrees and thus appear in a particular way because the specific user has visited it earlier (e.g. by means of cookies or of information that has been typed in).

Chapter 4

Archiving the dynamics of the Internet?

We can see a number of media-specific dynamics in the Internet, linked to sender, text and recipient, respectively: the dynamic of updating and proliferation, the dynamic of movement and complexity, and the dynamic of equipment and actions.[1] How are we to archive as many as possible of these dynamic aspects? Do we proceed with the aid of the document, the monument, or the imprint?

Document, monument or imprint?

Inasmuch as the Internet is a medium, it would be natural to think of its archiving as a monument or imprint, analogous to the archiving of newspapers, photos, films, radio, TV, etc. But as we have seen, the dynamics of the Internet are so fundamentally and qualitatively different from the circulation in the broadest sense of material/immaterial objects, which can simply be taken out of circulation and stored, on the whole unchanged, that it necessarily more closely resembles the dynamic character of nature and society, so that some form or other of the document seems to be the most obvious choice, taking into consid-

1. To this is added what you might call the general dynamics of the Internet: the fact that its forms, as regards sender, text, and recipient, are rapidly and incessantly changing: pure html pages have been supplemented by frames and cms; text first by graphics/images/sound, then by moving pictures and later by streaming, and finally, identical html pages for everyone have been supplemented by cookies and personalised pages (cf. above: that the dynamic characteristics are time-bound). Therefore, constant, general 'surveillance' of the Internet would be required if one wished to archive it on a continuing basis.

eration the high degree of complexity and changeableness of the Internet.

A document of the Internet

But is this a question of creating a document in the same sense as mentioned earlier, i.e. a document *about* the object? Both yes and no, and this is where the distinctive archival form of the Internet appears most clearly.

Archiving the Internet resembles the document: firstly, because it involves an element of *representation* of the object, insofar as the storage involves some form of 'transformation', and secondly, because the actual process of stabilisation involves a relatively high degree of *subjective involvement*.

But archiving the Internet is basically different from the document about the object, insofar as it in itself is a medium from the outset and therefore need not first be mediated/symbolised as the concrete object of analysis. It would thus be more correct to say that archiving the Internet does not create documents *about* the Internet, *but documents of the Internet*. As the object of analysis, the Internet constitutes a raw material which is already mediated, and in its archiving a new document *of it* is created. When documents of the Internet are created, there is thus a kind of remediation in the same medium: the already mediated forms the parent material for a new mediation.

This leads us to two basically different but equally possible ways of stabilising and saving the Internet, namely in the shape of *documents about* and *documents of*:

- On the one hand, we can create documents about the Internet, exactly as if it were any other 'piece of nature or society', i.e. observe the activity on the Internet and with the aid of quantitative or qualitative methods create other, new documents about the Internet that are fundamentally different from it,
- On the other hand, the actual archiving of the Internet, where we — within the limitations set by the dynamics of the Internet — attempt to create documents of the already mediated raw material available.

We can thus conclude that the Internet, with its dynamic characteristics, holds

a high degree of complexity and changeability in all its dimensions, so that it cannot be archived as monument or imprint, but only as document, either as document about the Internet or as a quite specific type of document: a document *of* the Internet.

The need for considerations of method

Apart from the appreciable consequences of Internet dynamics for the archived material (cf. earlier on the dynamic of updating), it is also worth noting that the actual archiving of the Internet ('document of') *is* still a document, and therefore has, as mentioned, the characteristics of the document, inasmuch as there is a certain degree of representation and subjective involvement. An important consequence of this is that archiving of the Internet should be accompanied by a number of deliberations as to method.

Firstly, what one wishes to examine in a later analysis should to a certain degree already be anticipated in archiving, since the concrete object of analysis not only exists (as monument and imprint), but to some extent is only created in and by the archiving. Therefore, methodical deliberations on *why* this document has been created should be an integrated element in its archiving.

Secondly, an archiving of the Internet should be accompanied by methodical considerations of *how* the document of the Internet has been created, just as when a document about the object of an analysis is created.

The two questions are also inseparable, inasmuch as the purpose of the following analysis ('why?') is decisive for how archiving of the Internet and its dynamics is to be attempted ('how?'). Therefore, the specific analytical purpose of the archived material (if known previous to archiving) should be taken into consideration as much as possible before and during archiving.[1]

---

1. This, by the way, is where the difference between what we began by calling micro and macro archiving becomes clear, in that micro archiving is usually based on a *specific* analytical purpose, while macro archiving is not aimed at any definite analytical need, but on the contrary, in principle archives 'everything'. However, micro archiving can be non-specific — e.g. if the purpose of archiving is either not sufficiently clearly defined — or not defined at all previous

It is difficult to say anything in general on why documents of the Internet are created — that depends on the specific analytical purpose — and for that reason the main stress in the following will be on the question of how the document of the Internet is created, with the discussion taking two paths: a test of archiving software and a discussion of the elements that can be included in an archiving strategy.

to archiving (see the section on 'purposelessness', pp. 58-60), and in contrast, macro archiving can also have a specific analytical goal such as on-demand archiving by a major archival institution.

Chapter 5

Test of archiving software

A comprehensive test of archiving software was carried out in connection with this book. The main question for the test was: how much of a website is it possible to archive with the micro archiving software available today?

Focus is placed on the Internet dynamics that are linked to the textual elements of expression (and to some extent recipient-specific dynamics), primarily because this book mainly concentrates on micro archiving.[1] Thus we do not aim to focus on the software and the principels of construction underlying or 'behind' the image on the screen, and which are linked to choices made by the sender (pure html pages, cms, frames, javascript, etc.), but rather on what one, as an ordinary user, can call forth on the screen, and how much of it can be saved. So the approach is 'wisiwiw': what I see is what I want.

A completely archived website?

As pointed out in chapter 3, the textual elements of the website are characterised by a dynamic of movement, by which is meant the movement manifesting itself *between* and *in* the individual elements of the website's structure (cf. appendix 1). So a completely archived website will contain:

- the entire structure and all its elements
- all actual and possible movements between and in the elements of the structure

If archiving software that can fulfil these requirements is available, the problem is solved. If not, an archiving strategy must be drawn up which — based on the

1. Regarding tests of software for macro archiving, see the results shown in Brügger et al. 2003: 31-44.

purpose of the analysis, if it has been defined — aims to bring the archived material as close as possible to the completely archived website.

Types of archiving software

In the test and in the book in general, we distinguish between four general types of software which can archive the Internet by:

- archiving entire websites (or parts of them)
- archiving individual webpages in a static form
- making individual screen shots
- filming/recording screen and/or speaker activity[1]

The last two types of software as such are not related specifically to Internet archiving, but are used to save any sort of activity that takes place on the computer screen (whether static image or film) or on the speakers, regardless of whether what takes place has anything to do with the Internet. Thus the two first actions are based on direct contact to the Internet, which is not the case with the two last.

Prerequisites and results

Since the subject at present is micro archiving, the choice of archiving software to be tested will be based on the following general conditions:

- The purchase price of the software must be as low as possible in order to ensure that as many as possible can make use of the test results
- Several platforms/operative systems are to be involved, in order to ensure a broad range of hard- and software in the test.
- The software must be as new as is feasible, in order to increase the likeli-

1. Streamed sound can also be archived with software dedicated exclusively to recording sound coming out of the computer speakers (e.g. WireTap, (for Mac), which can be downloaded free of charge from http://www.ambrosiasw.com/utilities/freebies). However, for these purposes archiving of streamed sound has only been tested as an element in screen-recording programmes.

hood of being able to archive the newest forms of expression.

Based on these criteria, seventeen programmes were chosen for testing. These programmes were primarily tested with regard to functionality and speed (i.e. the programme's ability to archive the various elements of which a website is composed, including the recommended settings and parameters, as well as how quickly the programme is able to archive a website and how much space the archived material occupies). For the test we used computer equipment comparable to what is typically available to researchers or students (see also the detailed account of the above-mentioned points in Thomasen 2004).

The chief result of the test was that none of the programmes tested was capable of completely archiving a website (i.e. archive the entire structure and all its elements as well as all actual and possible movements between and in the elements of the structure).[1] For a more detailed presentation of the test conclusions and results, along with recommendations for using the individual programmes; see http://cfi.imv.au.dk/eng/pub/webarc.[2]

1. This is actually not a new observation, but simply confirms earlier tests; see also Brügger et al. 2003: 39-40. The test carried out here is, however, more detailed in its focus on elements of content and expression, as well as also testing several different programmes.
2. In addition, one of the general test conclusions of more practical archiving interest was that it is strongly recommended to allow time to carry out one or more test archiving processes prior to the final process.

Chapter 6

Elements of an archiving strategy

Since none of the programmes tested can completely archive a website, the next step will be to sketch some possible archiving strategies, aimed at bringing the archived material as close as possible to a completely archived website. A strategy for archiving should be laid down, based on the purpose of the analysis, if defined, and also based on the following two conditions:

- an evaluation of how each of the main types of archiving software can best be used (and possibly combined),
- an evaluation of the extent to which it is possible to effect the strategy with the aid of the means available (software, hardware, platform, the quality of the archiving software, finances, time, etc.).

Laying down the strategy thus means answering the following three general and related questions:

- What is the *analytic purpose*?
- Which type(s) of *archiving software* should I choose in order to achieve my aim? And how should I combine them if I choose more than one?
- Can the *means available* comply with my purpose?

In a book such as this it is difficult to lay down concrete, detailed archiving strategies, since the number of existing websites as well as possible analytic goals is, if not infinite, certainly extremely large and varied. So the first strategic question will only be touched on in passing, mostly in the form of brief examples. In contrast, the second strategic question will be in focus, namely in so far as the elements of which an archiving strategy could be composed will be sketched, after which it is up to the individual to decide what building blocks to

use and how to use them in each case.[1]

Building blocks and variables

It will be useful to return to the four main types of archiving software mentioned in connection with the archiving test (archiving entire websites, individual webpages in a static form, individual screen shots, and screen recording), in that they each seem suited to specific archiving purposes. These types compose our building blocks. But in addition there are three important variables that we know each of the building blocks must consider, and they concern, in a broad sense, *space*, *time* and *montage*. Let us first regard the building blocks alone, then the variables, and, finally, how the two are related.

Building blocks: types of archiving software

The four main types of archiving software are each suited to specific purposes. If, for instance, the focus in the subsequent analysis of the archived website is to be on the website's structure (both single pages and several pages), on its elements and the ability to move freely between them, or on the movement of the elements when their activation does not require continuous online presence, then the first type of software can be used: software for archiving entire websites (or parts of them, including single pages).

If the analysis is to focus on the individual webpages, without any type of movement being possible, whether in or between elements, then one of the next two types can be used: software for archiving individual webpages in a static form, or screen shots.

Finally, if the focus of the analysis is to be on individual elements, where activation of the elements' movements requires continuous online presence, but where it is not important to be able to move freely within the structure of the

1. The third strategic question ('can the *means available* comply with my purpose?'), will not be discussed further, but will be covered in the test and its results. Finally, as suggested in the above, when choosing a strategy/strategies it is important to keep these very concrete conditions in mind.

archived website, then the last type of software — for screen recording — can be used.[1]

As we can see, each type of programme has certain shortcomings:

- the first cannot archive movement in the elements if the activation requires continuous online presence
- the next two cannot archive any form of movement, whether in or between elements,
- the last type cannot archive the ability to move freely within the structure of the website, in that it is always necessary to follow the direction of the original screen recording.

Variables: space, time, montage

As previously mentioned, there are three kinds of variables that must be considered in connexion with each of the four types of archiving software. They will first be treated in a somewhat abstract manner, then more concretely when subsequently combined with the building blocks.

The first variable concerns what one might in a broad sense call the *spatial extent* of the website, and there are two things to be considered here:

- How large is the website to be archived? And is an entire website or part of a larger website to be archived?
- How large will the result of the archiving be — In other words, how large will the archived file(s) be?

---

1. Depending on the degree of user intervention required in connection with continuous online presence as a prerequisite for activating movement, what is archived using screen recording will range from something always the same in all cases (such as archiving a streamed news item) to something varying over a limited set of possibilities (such as the choice of programme/camera angle), to something that is a specific example of innumerable possible activations of the movable element (such as actually participating in a game or a chat; cf. the section on text, p. 26). The more specific and open the mobile activity is, the more we are dealing with an example, and in screen recording with this type of element, one can never archive the *possibility* of playing or chatting, but only examples of games actually played and chats that have taken place.

The reason we must consider the size of the website is that the larger and more complex a website is, the greater the likelihood of error in archiving (both software and human), and the more difficult it is to find any errors and begin archiving again; and conversely, the smaller the website, the less likelihood of error. This means that if we are speaking of a large and/or complex website, we can be forced to split archiving up into smaller elements, which raises two new questions:

- How large should what we might call the *archival element* — i.e. the parts the website is divided into — be? And according to what criteria should division be determined?
- What *direction* should archiving of the archival elements take? And, again, according to what criteria should the direction be determined?

The second variable concerns what we might call the *temporal extent* of the website, and here there are two questions to be considered:

- How long will it take to archive a website of the size in question?
- How long is it necessary for a person to be present during archiving?

Again we must consider the size of the website, but now for reasons of time, connected with the risk constituted by the dynamic of updating, since it tends to increase along with the size of the website: the larger the website, the more difficult it will be to minimise the problems posed by the dynamic of updating, and conversely: the smaller the website, the better the feasibility (although the risk can never be eliminated).[1] So therefore one can be forced to break up archiving of the website for reasons of time if it is a question of a large website, so that the temporal aspect also plays a part in deciding the archival element and direction.

The third variable concerns *montage*, and here we must consider how any archival elements are to be assembled again, so that the archive's website, both in spatial and temporal regard, resembles as closely as possible the website that

1. The dynamics of updating are not only a question of size. A small website can very well have a high rate of change, and a larger site a low rate of change, but other things being equal,

has been archived.[1] In practice, montage takes place via the manner in which the archived material is stored and accessed, usually by the individual archival elements being opened, each in its own window (so that it will be necessary to consider how this is to take place before beginning archiving; cf. the 'Step-by step-guide' in appendix 2).

Building blocks and variables

More precisely, how are the three variables, space, time, and montage distinguished, when archiving is to be carried out using each of the software types mentioned?

1) Software for archiving entire websites

If we begin by looking at the spatial extent of the website, any problems regarding size will be easily assessed if it is a small, uncomplicated website. If, however, one wishes to archive a large and/or complicated website (or part of one) using software for archiving entire websites, size may pose a problem, and one would therefore as a matter of course need to consider what is to be the archival element and the archiving direction. In order to be able to do so, one should begin by forming an overview of the entire website, a process that should result in a *diagram of its structure*.[2] The diagram should be drawn up as near

---

there is generally less risk in the case of a smaller website.

1. The main, general goal will be for the website in the archive to approach the website being archived, but in certain analytical connections it may be more practical to assemble the archival elements in other ways in order to better make allowance for the purpose of the analysis (which might then also be reflected in the division into archival elements).

2. This site diagram should not only form the basis for determining what is to be archived in each archival element, but to just as great a degree what is *not* to be archived. With the aim of delimitating each archival element as clearly and precisely as possible, it is an advantage to specify URLs that are not to be archived, which may be especially useful on websites where there are many links between the individual areas. If in such cases one requests, for instance, the archiving of a sub-directory, along with the links from it, large parts of such a website will be archived if one does not state precisely what not to include (a number of archiving soft-

the time of archiving as possible (to avoid changes to the website structure before archiving), and could, for instance, be inspired by the menu structure of the website (main menu, thematic menu items, etc.), its site map, or something similar, but since it is to be created with a view to archiving the website, its structure will not always necessarily coincide with the technical and/or editorial structure of the website.[1] The most suitable archival elements are chosen on the basis of this site diagram (e.g. front page, sub-areas 1,2,3, etc., further sub-areas 1.1, 1.2, 2.1, 2.2, 3.1, 3.2, etc.), and the direction of archiving is determined (e.g. from top down (front page, area 1, area 1.1, area 1.1.1, etc.) or based on levels (front page, area 1,2,3, area 1.1, 2.1, 3.1, etc.)). What is more precisely understood by 'most suitable' naturally depends on the individual website, and determining archival elements and direction will obviously not always be simple and straightforward. In addition, one must also be prepared for the possibility of being forced to alter both the archival element and direction during the process, if the archived material turns out to be defective.[2] Finally, the degree of detail of the site diagram will naturally depend on the website in question, so that the process is one of trial and error.

Based on these general guidance lines for the spatial context, the archiving of a large/complex website, using software for archiving an entire website, could proceed as follows:

- a site diagram is drawn up and the necessary logins are acquired

---

ware programmes allow for listing the URLs that are not to be archived).

In addition to this, when drawing up of the site diagram, one should ascertain whether it is necessary to login to the website or any part of it. If so, the login must be made available before beginning archiving.

1. This type of site diagram will be vital to the archiving of a large website as a whole, but it will also be applicable if one is only archiving part of a large website, as it offers an overview of the (immediate) context of the archived material.

2. Apart from allowing for an overview of the website's structure with an aim to choosing archival elements and direction before archiving, the site diagram can act as an archiving log during the process (and later be used as general documentation as well as possible certification for sources of error).

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

It is not recommended that archival elements and direction be altered too frequently or too drastically, as one may then be forced to start over every time.

Finally, in the spatial context it must be taken into consideration how much

1. Validation compared to the website in existence on the Internet is most easily done by opening two windows of the same size, placing them on top of each other, then paging down, shifting between them, paging down, shifting, and so on. However, strict attention should be paid to two possible sources of error. Firstly, the archived website, often with no warning, can obtain material from the active Internet if you are online — which you are, since you are comparing the archived material to that on the Internet. Secondly, even if you go offline, it will be possible for the archived website to obtain material that is not actually in the archived material, but which is temporarily on the computer in a so-called cache (the area on the hard drive and/or memory where recently visited websites 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.

space the archived material will occupy — in other words, how large the archived material's file(s) will be. The size of the files is, of course, dependent on the size of the website, but in general it can be said that the size of files in this type of software is typically measured in megabytes.

If we now look at the size of the website in a temporal context, the problems related to the dynamic of updating are minor if it is a matter of a small website. If, however, one is to archive a large and/or complex website (or subsite), which it is necessary to divide into its parts, then the risk of inconsistencies based on the temporal aspect increases. When determining the most suitable archival elements and not least the archiving direction, the temporal aspect may need to be considered, insofar as certain archival elements may be presumed to have a high rate of change, such as front pages or news pages, while others can be presumed to have lower rates of change, such as areas acting as archives. If a website has several areas with high rates of change, and these have some form of semantic connection, it can be an advantage to archive them at the smallest possible intervals of time, even though this is not logical based on the structure. For instance, in the case of a website with a front page with brief news items/headlines, referring to one or more underlying pages with the actual news texts, one could consider archiving the front page and these underlying pages sequentially, saving areas such as those with archived material until later. One possible way of minimising the temporal distance between two or more archival elements is to begin the archiving of each of them at the same time, or at very short intervals. However, it is a prerequisite in this case that this is possible with the software in use (which is usually the case), and also that there is enough memory and storage capacity for this to be done without any problem.

Thus, in the process of archiving sketched in the above, it is possible for validation to also have a temporal aspect, in that in many cases it is a question not only of comparing with what was/is on the Internet, but also with what was earlier archived, with the aim of clarifying to what extent the temporal aspects 'fit' (for instance: are the news items referred to on the front page actually to

be found on the archived underlying pages? In most cases, however, there is not much that can be done if they are not, except perhaps to further minimise the interval between the two).

When the temporal aspect is involved, the process of archiving will be as follows:

- a site diagram is drawn up and the necessary logins are acquired
- 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
- archival element 2 is archived
- 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.

Finally, the question must be considered of whether it is necessary for a person to be present during archiving. Theoretically this is not necessary when using this type of programme, but occasionally it is necessary to be present to some extent, in that archiving sometimes stops, e.g. if it cannot proceed because of a missing password for another server (where the operator must check 'no'); archiving stops for no apparent reason, or the computer simply crashes. In any case, a presence is needed in order to validate the archived material, both in

relation to the website on the Internet as well as in relation to the archived material, and the greater the number of parallel archiving processes that are active, the greater the need for a constant presence for validation.

With the above question of validation in relation to what is found in the archive, we move de facto into the area of the last variable — montage — in as far as it is a matter of comparing two archival elements. The most important aspect of montage is that already when determining archival elements and the archiving direction one must as far as possible anticipate their montage after archiving — i.e. whether, and if relevant, how, the archival elements can be reassembled so that the archive's website approaches the website that was archived as closely as possible with regard to both spatial and temporal aspects.

With this type of software it will now often be seen that the spatial demarcation of archival elements that has presumably taken place (e.g. a smaller area of a website based on an URL, possibly with a certain number of specified subareas) does not always agree with what has actually been archived. Often more (or less) has been included, which may mean that the same material can have been archived twice (but may still differ on certain points) or there may be material missing. The limits of the archival elements will thus often turn out to be border *areas* rather than border*lines,* which can make later montage difficult. Added to this are the montage problems presented by the dynamic of updating. The possibly varying temporal aspect of the archival elements will thus often mean that it appears that one does not archive a website at a given *point of time*, but rather in a given *period of time*, in which changes can have taken place.

So when determining the archival element, two opposing forms of logic regarding space and time apply from the point of view of montage: if one uses small archival elements, the entire process will take only a short time, which minimises the risk involved in the dynamic of updating, but at the same time we increase the number of archival elements and thus the number of 'border areas' that must fit together in montage; and conversely: if one uses large archival elements, the process will take longer, which increases the risk involved in the

updating dynamic, but at the same time the number of archival elements decreases, and with them the number of 'border areas' that must fit together in the montage.

2) Software for archiving individual webpages in a static form or for screenshots
Software for archiving individual webpages in a static form or for screenshots can be regarded, as in the above, as one, since the result in both cases is a static image with no possibility of any kind of movement. Apart from the fact that these types of software are not capable of archiving any form of movement, neither can they archive an entire website in what might be termed an unbroken state. Since it is a case of archiving or photographing single pages, the website, regardless of size, is always already broken into archival elements. However, the limits of the archival element are not defined in the same way in the two types. In the first type (archiving single pages in a static form) the limits of the archival element are defined by the archiving software, in that the archival element is always a specific URL, while the outer limits of the archival element in the other type (screenshots) consists of the actual surface of the screen, so that the archival element is not given to begin with, but rather is created with the aid of any cropping of the screen image that may take place. A consequence of this is that although software for archiving single pages in a static form can only archive *one* web page, it does include what cannot be seen on the screen (so the archival element is (often) larger than what can be seen on the screen); while a screenshot, although it can only archive what can be seen on the screen (thus not the parts of the pages that are not visible on the screen), can archive *several* web pages at the same time (in the shape of open windows), that is, as many as it is possible to have open on the screen, depending on its size.

So the size of the website is not important for the size and demarcation of the archival elements — they are mostly given in advance — but with this type of software it can in any case be an advantage to create an overview of the website, so that especially the archiving direction can be determined. However,

following this, the process of archiving will be less complicated, since validation will either be very simple or completely unnecessary, so that it rarely gives rise to corrections in the direction (validation is only necessary when archiving a single web page, and is only a question of whether the entire web page and all of its elements are there or not, and if not, it cannot be remedied by changing direction, but only by archiving the web page again (besides, one is often given a preview of the archived material before saving it); validation is not necessary with screenshots, since the image does not have direct 'contact' with the Internet (the screen image is what is included in the screen shot)).

Archiving with this type of software can proceed as follows:

- a site diagram is prepared, and the necessary logins are acquired
- 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.

Finally, with regard to the spatial aspect, one must consider how much space the archived material takes up — i.e. the size of the file(s). In this type of programme the size of the individual file does not depend on the size of the website, but on the quality of the individual image (in the cases where it is possible to adjust it), and the size of the file will always be the same for all images, given the same image-quality; if one wishes to thoroughly record a website (or parts of one), it is the number of files that is decisive rather than the size of the individual file. Incidentally, the size of files in this type of software is usually measured in kilobytes.

If we look at the temporal aspect, then the dynamic of updating constitutes a minor problem with this type of archiving software, since there is a greater potential for minimising the temporal distance between rapidly changing web

pages/areas. This is because, firstly, the archival element is already given as an entity, which in a temporal respect is clearly limited to a point in time, and secondly, continual validation and possible adjustment are (on the whole) unnecessary, which allows for faster archiving, and finally, archiving is not dependent on the speed of the archiving software, but on how quickly the person doing the archiving can work. This last point then calls attention to another aspect of the question of time: that the presence of a person is necessary during the entire process, which influences the length of time needed for archiving.

Finally, with respect to montage, when archiving with this type of software, it will always be necessary to anticipate a subsequent montage (if more than one web page is archived), in that this type of archiving software, as mentioned, cannot archive an entire website, but only parts of one. However, montage is relatively simple, since all archival elements will have clear borderlines, and will always have been archived at a precise point in time. Screenshots may, though, lead to more complicated montage if there are two or more web pages open in each static image.

3) Software for screen recording

With regard to the type of software used to record what is taking place on the screen, one can relate to the website in two ways:

a) one can either move around in it — for instance, if one wishes to archive an entire website or areas of a website,
a) or one can refrain from moving around, for instance if one, in the same element and over a period of time, wishes to record movement requiring continuous online presence (such as streaming).[1]

Each of these gives rise to different kinds of considerations of space, time and montage, for which reason they will be treated separately in the following.

a) If one intends to archive an entire website by moving around in it and

1. Recording where one does not move around in the website can, of course, also be included in recording where one does.

recording it at the same time, the size of the website is important, not because errors in archiving can be generated by the archiving software (as they can by software for archiving entire websites), but because when using the recorded material later it can be difficult to find specific webpages if an entire website is recorded on one film. In contrast to software for archiving single web pages in a static form or for screen shots, this type of software can archive an entire website in what one might call an unbroken state, which, however, is also a problem, in that the only way in which one can navigate the film/the archived website, is by spooling.

As in software for archiving entire websites, it is therefore recommended that archiving be divided into archival elements which are archived in one direction (but in this case for reasons of convenience, not technical reasons). Here, however, the demarcation of the archival elements differs from the other two types, in that ultimately it is the person who does the archiving who defines the limits, independently of any aspects of the archived material and the archiving software. The archival element can therefore be clearly delimited, but the criteria for defining the limits can only be drawn up by the person doing the archiving. Here, too, it can be an advantage to form an overview of the website by drawing up a site diagram, both in order to define the archival element and the archiving direction (although the direction is seen on the film, it can be relevant to locate the beginning(s) of the recording, especially in larger websites/areas of websites). As with software for archiving individual web pages or screenshots, the archiving process will now be less complicated, since validation in relation to what is on the Internet is unnecessary, because the recorded material has no direct 'contact' with the Internet (the material being recorded is what is on film); any validation will simply be a kind of 'internal validation' — i.e. were the archival element and direction appropriate? — and will, of course, lead to any necessary adjustments.

Archiving movement around a website with this type of software can proceed as follows:

- a site diagram is drawn up and the necessary logins are acquired (remember

that the login will appear on the screen recording)
- 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.

With regard to file size, it depends on how much is recorded — i.e. the size of the website — as well as on the image quality of the recording. The file size will always be the same for all recordings, if the length and image quality correspond. File size in this type of software is usually measured in megabytes, and if the image quality is to be acceptable, we will quickly arrive at large files (one minute's filming will give a file of approx. 10 MB (with mono-sound, 15fps)).

As to the temporal aspect of recording an entire website, the risk arising from the dynamic of updating is the same as if archiving had been done using software for archiving an entire website. However, there is a greater potential for minimising a distance in time between rapidly changing web pages or areas: one can decide the size of the archival element, and it can be clearly delimited; continual validation and adjustment are (on the whole) unnecessary, and archiving is not dependent on the speed of the archiving software, but on how quickly the person doing the archiving can operate.

Finally, as part of the temporal aspect it should be noted that the presence of a person is required during the archiving process, which influences the length of time needed for archiving.

With regard to montage, during archiving it will be necessary to anticipate a subsequent montage if archiving is carried out in archival elements — i.e. short lengths of film. Montage will, however, often be less complicated than when using software for archiving entire websites, since one can be certain that the archival elements and direction that one settles on are also those that are actually archived, just as they will always have clearly defined borderlines and will always be archived at a precise point in time. However, recording an entire

website can be a less appropriate way of archiving a website, partly because one loses the potential of later being able to move freely in the archived website, partly because the process is weighty (file size increases rapidly and a constant presence is required).

On the other hand, screen recording can be more applicable for archiving movement in an individual element where the archiving requires continuous online presence — i.e. usage with no movement around the website.

b) In contrast to recording a website by moving around in it, the size of the website is unimportant when simply recording a movable element (a single page, a single window, part of a page, etc.). The archival element coincides with the archived material, and there is no archiving direction. However, recording the (immediate) context of the element might be considered, if it may in any way be relevant to a subsequent analysis (if, for instance, one wishes to archive a single news item, it can be relevant to also film the web page associated with it; this might be done as a 'lead in' to the news item to be filmed). A site diagram is thus not as necessary for this type of archiving, but can be drawn up, especially if it is important to have an overview of the context of the element. Validation in relation to the material on the Internet is unnecessary, since the recorded material does not have direct 'contact' with the Internet (what is recorded is what is on the film).

The archiving of motion in a single element can proceed as follows:

- a diagram of the site or the closest context of the archived element may be drawn up and the necessary logins acquired (note that the login will appear on the screen recording)
- 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).

Questions of file size are basically the same as when recording an entire website (for instance with respect to image quality), except that it is not the spatial extent of the filmed material, but its extent in time that determines file size. If long streaming sequences are to be recorded, file size will quickly increase.

Problems related to the dynamic of updating play no role in the temporal aspect, since all updates are included (though it can be important if the context is archived in a separate film). Since there is no question of recording motion on the website, the presence of a person during archiving is theoretically not necessary, but as it will often be a case of small screen recordings because the files rapidly become very large, a person will usually be present during the entire archiving process. With respect to montage, this is only necessary if the context of the movable element is recorded on a separate film.

Combined forms and purposelessness

Up until this point the elements of an archiving strategy have been discussed in what one might call their 'pure' form, with precisely one type of software filling the need for archiving presented by the final analysis. However, there will always be two further possibilities: firstly, that the need for archiving cannot be filled by only one type of software, so that several types must be involved, and secondly, that the purpose of archiving is not sufficiently clearly defined or is completely unknown prior to archiving, complicating the choice of the most relevant type(s) of software. These two situations will be discussed in the following.

Combined forms

The archiving of a website will often require the use of some combination of the four types of software, which in each case raises new questions concerning space, time and montage. We will briefly treat three of the most important forms: *documentation*, *exemplification* and *contextualisation* (this is not an exhaustive list; there are of course, other possibilities). The conditions earlier mentioned with regard to each type of software still apply, for which reason the following focuses only on the areas where the combined forms differ from them.[1]

1. In all three of the combined forms treated, it is recommended that the two archiving proc-

1) Documentation

If it is important to the subsequent analysis of a website that all its elements of expression are actually archived, and moreover, are archived as they actually appeared and were placed; or if during the archiving of the structure one becomes aware of changes in a web page that for some reason or another are important, then it can be advantageous to combine the use of software for archiving entire websites with either static or dynamic *documentation* in the form of static images or a screen recording (here, screen recording is not used to archive movable elements needing continuous online presence). In this way, elements that may later turn out to be lacking in the archived structure will be documented. In illustration, one example among many: the question of documentation can arise if one wishes to analyse the design history of websites and needs to be certain of page construction, and so on.

Unfortunately, often the elements of a website are not all archived when software for archiving entire websites is used, which can be the result of conditions specific to the programme (the archiving software is not set up correctly; for some unknown reason certain elements are not archived; there is not enough memory, etc.; cf. the test results), or the result of the dynamic of updating mentioned earlier. Any errors or omissions should appear during validation, unless there were doubts from the beginning as to the quality of the hard- and software used.

With smaller websites one may choose to document everything, while with larger websites certain areas must be chosen, based, for instance, on the areas that are central to the later analysis, or on the fact that certain web pages in general are especially important. This might be, with regard to space, web pages that compose navigational 'crossroads', linking many pages (front page, the front pages of smaller areas…) or often, with regard to time, web pages with high updating frequencies. Documentation using either static images made

---

esses 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).

at short intervals or recording, along with archiving using software for archiving entire websites, thus constitutes a way in which one can archive the changes taking place while the structure is being archived.

Finally, it must be decided whether to archive using static images or recording. Static images require many single archiving processes, but they are to a certain degree easier to navigate in subsequently, while recording is a rapid method since one needs only to move to the material to be documented, after which it is archived, but the recording can, as earlier mentioned, subsequently be difficult to navigate in, especially if it is long. And, of course, documentation using static images results in many small files, while recording results in fewer, but larger files. Regardless of whether documentation is done using static images or recording, it is important to indicate in the site diagram where it has taken place.

Things become more complicated with regard to montage, and this applies to all combined forms, since one must now not simply assemble archival elements from the same type of archiving software, but make elements in each of the two types of archiving software fit together, as well as making elements from two different types of software fit together. So this is not a question of two parallel montages (each with the problems of time and space earlier described), but also of a *third montage,* where two different types of archiving of 'the same thing' must subsequently be combined.

If a large and/or complicated website is involved, this third montage can be complicated, especially if screen recording is used, in that the recording can move crosswise in the structure, in contrast to static images, which can always be localised as a single point in the structure. So in spite of the fact that with screen recording the archival element is clearly delimited, its limitations do not necessarily compare to anything easily localised in the structure, so that it can be difficult to determine the recording position in relation to it. In order to avoid this problem, one can either use static images exclusively or be extremely painstaking in placing the screen recording in the site diagram in relation to the archived structure.

2) Exemplification

If in a subsequent analysis of a website it is important to be able to give examples of specific mobile elements or types of elements requiring continuous online presence, without making them the object of a thorough analysis, the use of software for archiving entire websites can be combined with an *exemplification* of the mobile elements in the shape of a screen recording. This could be the case if one wanted to analyse a news website containing news items in the form of streamed sound or image clips as it appeared on a given day; then a screen film of one of these news clips could exemplify how the news item functioned alone and together with the remainder of the web page.

So exemplification is a combined form that does not need to compensate for errors in the software or the dynamic of updating, but rather for one of the lacks in a certain type of software mentioned earlier, namely, that software for archiving entire websites cannot archive motion in the elements if this requires continuous online presence. The size of the website as such is not important for exemplification, if, of course, one knows the position in the structure of the movable element which is to be exemplified (which is comparable to knowing what one wants to document when documenting). However, the number of elements to be exemplified can constitute a problem.

With regard to montage: exemplification also involves three montages, but the problem regarding screen recording that was pointed out in connection with documentation does not apply here, in that recording is not done by moving around the website. On the contrary, a single movable element — a single point — is recorded, so that the screen film, from the point of view of montage, becomes identical to the static image and thus more easily localised in relation to structure.

3) Contextualisation

If it is important that a subsequent analysis of a website allows for a thorough analysis of the website as a seamless textual system of expression — in other words as structure, elements, ability to move between the elements and the

ability to move in all elements, including those where activation requires continuous online presence — then one can use a combination of software for archiving entire websites and software for screen recording. In this way the materials archived with the two types of programme constitute each other's *contexts*. This combined form resembles exemplification, but the two types of archiving software are equally important, and also it will usually be necessary to use screen recording to archive more material if the archived material is to serve as the object of a thorough-going analysis. This can be the form to use if, for instance, one wishes to analyse a radio/TV station's teenage website during an entire day, with everything it contains of running updates, news items, AV streaming, webcams, clickable maps, quizzes, chats, games, and so on.

Like exemplification, contextualisation need not compensate for errors in archiving software or the dynamic of updating, but it does need to compensate for lacks in software for archiving entire websites and screen-recording software — i.e. for the fact that the type of programme in the first case is unable to archive movement in the elements if its activation requires constant online presence, and in the second case cannot activate the ability to move freely in the structure of the website.

But the archived material can never be truly seamless, as is the active website, in that the deficiencies in each of the two types of programmes remain, in spite of being partially compensated for by combining the programmes. However, the combined form of contextualisation is the closest we can come to eliminating what has been called the dynamic of complexity. As in exemplification, the size of the website as such does not play any particular role in contextualisation if we know the position in the structure of the movable element to be exemplified.

But since the goal is archiving with a view to a more thoroughgoing analysis, this combined form will often require large amounts of material to be archived, usually with the aid of screen recording. This can mean that the subsequent three montages (which are basically identical to montage in exemplification) can be made more complicated for purely numerical reasons (again, the site

diagram is important).

Purposelessness

As earlier mentioned, it often happens that the purpose of the archiving is either not sufficiently clearly established or actually not known prior to archiving, and it will therefore be difficult to choose the most relevant type(s) of software and possible combinations.

Often the purpose of the archiving is not sufficiently clearly established if one or more websites have been chosen for analysis before determining the analytic focus, theoretical/methodical approach, etc. There can be many reasons for having no idea of the purpose of archiving while wanting a copy of a website as it appeared at a given time: perhaps it simply belongs to an area of interest or research, and is felt to be a possible subject for later analysis, perhaps there is a suspicion that the entire website, or some of the material on it, may suddenly disappear (e.g. websites that are illegal for whatever reason, and are therefore suddenly closed down or moved to other locations on the Internet, or websites with incriminating material, since removed), or perhaps the website is related to a sudden event, making it interesting to archive for that reason alone (such as 9/11, wars, natural disasters, a sudden election, etc.).[1]

The absence of a clearly defined purpose for archiving means that the guidelines for choosing a strategy are lacking, and it is therefore necessary to choose strategy more or less blindly, meaning that it should be considered that the archived material may not be suitable for use in subsequent analysis. However, in an attempt to ensure that what is archived is of at least some value, we can point to one problem in general and at the same time formulate four 'imprecise' strategies.

The movable elements requiring continuous online presence, such as

1. As earlier mentioned, the lack of a clearly defined purpose makes purposeless micro archiving resemble macro archiving, which, after all, is characterised to a certain degree by the fact that in archiving the Internet as part of our cultural heritage we cannot know what analytic needs it will be the subject of at present and in the future.

streamed sound and certain games requiring user intervention (they must be started and may require interaction), constitute a very important problem. If this type of element is to be archived, the problem is to find out *whether* they are on a website, and if so, *where*. If there is time to look for them, and if they have been found, one can proceed either by screen recording alone or use one of the combined forms of which it is a constituent, and, for instance, as already mentioned, carry out two archiving processes at the same time.

This problem leads to the formulation of four 'imprecise' strategies, where the two related circumstances to be considered are the *size* of the website and the *time* available for archiving.

If it is a *small* website and there is *little time* available, one should attempt to archive it by using either software for archiving entire websites or static images. If possible, it is advisable to obtain an overview of the website (the lack of time speaks against this, the small size for), and from this it may be possible to document central web pages. There is not much time to search for and record elements that require user intervention and online presence, but on the other hand, the small size means that there are not many places to search. There will most probably not be time to check the quality of the archived material.

If it is a *small website* and there is *a lot of time*, archiving can best be done with the aid of software for archiving entire websites, and since there is more time, it will be possible to form a better overview and to document central web pages, as well as to consider the two remaining combined forms. There will now be more time to search a smaller area for elements requiring user intervention and online presence, and there will likewise be time to check the quality of the archived material.

If, however, it is a *large website* and there is *little available time* the task is more difficult. Archiving will probably best be done with the aid of software for archiving entire websites, but it will probably be impossible to form an overview, so that any documentation or the archiving of elements requiring online presence will be difficult. And there will almost certainly be no time to check the quality of the archived material.

Finally, if it is a *large website* and there is *a lot of time* available, archiving will probably best be done by choosing software for archiving entire websites, and it is now easier to form an overview (although size and time are counteractive) so that documentation, as well as one of the remaining combined forms, is also possible. There may be time to search for and record elements requiring online presence, but the size of the website is counteractive here. It will also be easier to check the quality of the archived material.

In conclusion, one can say that the more time available, the greater the possibility that purposelessly archived material can fulfil a future purpose.

Chapter 7

Representation and subjective involvement

As mentioned earlier, an archived version of the Internet posesses the characteristics of the document, insofar as there is a certain degree of representation and subjective involvement. But how has it been shown in practice that the Internet cannot be stacked like a stack of newspapers, photographs or film, or cannot be recorded by pressing a button?

Firstly, a number of choices must necessarily be made within the framework set by each type of archiving software. In the case of software for archiving entire websites, certain questions must be addressed: is the entire website to be archived? How many levels is the archiving to cover? Are photos, sound and moving images to be included? Is material to be collected from other servers? — and so on and so forth, while with screen-recording software, these questions are important: How many images per second are to be included? Is sound to be included? Are mouse movements to be included? Etc.

Secondly, a number of choices must be made with respect to the archival elements and archiving direction. When using archiving software for entire websites, and for archiving individual pages in a static form, the framework for these choices is composed of conditions on the Internet (one or more specific URLs), as well as the manner in which these can be converted into archival elements and connected in an archiving direction.

The situation is slightly different when using software for screenshots or screen recording. In both cases it is necessary to choose a section: the entire screen, an area, or an individual window. These are on the whole the only possibilities when making screenshots, but screen-recording makes further demands, in that the choice of what the recording is to treat is entirely up to

the person recording. So the archival element can be freely defined within the limits of the computer screen's image as such, both on the surface and in depth. The screen image — both the surface picture and that hidden in the depths, i.e. the individual window plus underlying windows, desktop, etc. — lie like an open landscape for screen recording. In contrast to the filming of an 'ordinary' landscape, this type of remediated screen-recording is characterised by having only one point of view: one cannot change the viewpoint — i.e. see what is taking place on the screen from a different angle — so this is a kind of 'flat' fixed point of view, where physical movements of the camera have no influence, since 'the camera' is one with the screen. But regardless of this, in screen recording we are nearing a kinematic approach to archiving, where the person doing the archiving becomes his or her own director by treading their own paths in the landscape of the screen.

And thirdly, a number of choices must often be made with respect to montage of the archived material. For instance, it may be a question of collecting archival elements made with the same or different types of archivng software, and the question is whether the result in each case is an 'original', that may not have been on the Internet? So the person doing the archiving becomes his or her own film editor.

In these concrete ways, the creation of a document of the Internet holds a certain degree of representation and subjective involvement. This, however, seems to be a basic condition for the stabilisation of the Internet as an object of study.

Bibliography

Brügger, Niels (2001). "The last page of the Internet?", in *Preserving the Present for the Future. Conference on strategies for the Internet. Proceedings*, Copenhagen: The Royal Library/Denmark's Electronic Research Library, pp. 43-53. Found online at: http://www.deflink.dk/upload/doc_filer/doc_alle/846_Trykt proceeding.pdf

Brügger, N. & Finnemann, N.O. (2001). "netarkivet.dk", in *Politologiske studier*, 4, 4, Copenhagen, pp. 66-73. Found online at: http://www.politologiske.dk/artikel07-ps12.htm

Brügger, N. & Carlsen, S.V., Christensen-Dalsgaard, B., Finnemann, N.O. Fønss-Jørgensen, E., Henriksen, B. von Hielmcrone, H. (2003). *Experiences and Conclusions from a Pilot Study: Web Archiving of the District and County Elections 2001. Final Report for The Pilot Project "netarkivet.dk"*, Copenhagen: netarkivet.dk, 60 p. Found online at: http://www.netarkivet.dk/rap/webark-final-rapport-2003.pdf

Christensen-Dalsgaard, Birte (2004). "Web Archive Activities in Denmark", in *RLG DigiNews*, 8, 3 (June 15), Department of Research, Cornell University Library, 6 p.: http://www.rlg.org/en/page.php?Page_ID=17661"\l "article0 (downloaded July 8 2004).

Ess, Charles and the AoIR ethics working committee (2002). "Ethical decision-making and Internet research: Recommendations from the aoir ethics working committee", approved by AoIR, November 27, 2002: http://www.aoir.org/reports/ethics.pdf

Finnemann, Niels Ole (2001a). ”Internet — a cultural heritage of our time”, in *Preserving the Present for the Future. Conference on strategies for the Internet. Proceedings*, Copenhagen: The Royal Library/Denmark’s Electronic Research Library pp. 31-42. Found online at: http://www.deflink.dk/upload/doc_filer/doc_alle/846_Trykt proceeding.pdf

Finnemann, Niels Ole (2001b). ”Arkivering af internettet”, in *Kritik*, 34, 151, Copenhagen, pp. 13-18.

Finnemann, Niels Ole (2005). *Internettet i mediehistorisk perspektiv*, Copenhagen: Samfundslitteratur, 331 p.

Merzeau, Louise (2003). ”Web en stock”, in *Cahiers de médiologie*, 16, Paris, pp. 159-160.

Preserving the Present for the Future. Conference on strategies for the Internet. Proceedings, Copenhagen: The Royal Library/Denmark’s Electronic Research Library, 145 p. Found online at: http://www.deflink.dk/upload/doc_filer/doc_alle/846_Trykt proceeding.pdf

Thomasen, Bo Hovgaard (2004): ”Tests of software and strategies for micro-archiving websites”, Århus: The Centre for Internet Research: http://cfi.imv.au.dk/eng/pub/webarc/testforudsaet_thomasen.pdf (downloaded November 2004)

IIPC, International Internet Preservation Consortium (2004). ”Press Release”, at http://www.netpreserve.org, May 5: http://www.netpreserve.org/press/pr20040505.php (downloaded Sept. 9 2004)

Appendix 1

Typology of movement in elements

		writing/print	image	sound
automated	inherent	▪ Scrolling texts, blinking/movable letters, etc.	▪ banner ads	▪ background sound ▪ banner ads
	online	▪ scrolling texts ▪ chat as reader (no login)	▪ banner ads	▪ background sound ▪ banner ads
user inter-vention	inherent	▪ archived chat ▪ mouse-over ▪ quiz ▪ clickable map	▪ non-streaming image (e.g. slide show, click-able map, quiz) ▪ games ▪ quiz ▪ clickable maps (w. zoom or activation of smaller elements) ▪ mouse-over	▪ non-streaming sound (activated in games, quiz-zes, etc.) ▪ mouse-over
	online	▪ chat as participant (or reader w. login) ▪ polls ▪ test-yourself	▪ streaming (both ar-chived and live) ▪ games	▪ streaming (both archived and live)

The forms named in the above table are not a complete list, only examples.

Appendix 2

Step-by-step guide to archiving a website

Prior to archiving

The analytic purpose ii
Existing archives ii
Supplying by the producer ii
Ethical and regulatory questions ii
Types of archiving software iii
Specific means of archiving vi
Site diagram vii
Anticipation of subsequent treatment vii

The archiving process

Archiving software for entire websites ix
Archiving software for archiving of individual webpages
in a static form or for screenshots x
Screen-recording software (with movement on the website) x
Screen-recording software (without movement on the website) x

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.

Archiving Websites.
General Considerations and Strategies

This book treats the micro archiving of websites, i.e. archiving by researchers, students or others without special technical knowledge who, using a standard computer, wish to save a website for further study. The phenomenon is discussed from the standpoint that Internet research must be able to stabilise and save the object of its analysis. However, the Internet is endowed with certain fundamental media-specific dynamics that make stabilisation difficult. Based on an account and discussion of these dynamics (linked as they are to sender, text and recipient) the following double conclusion is reached.

Firstly, unlike other well-known media, the Internet does not simply exist in a form suited to being archived, but rather is first formed as an object of study in the archiving, and it is formed differently depending on who does the archiving, when, and for what purpose. Secondly, this means that there is an element of subjective creation in the archived material, so that methodical deliberations are necessary — in other words, the answers to why and how the archived material has been created. These conclusions form the starting point for the last section of the book, which, based on comprehensive tests of archiving software, discusses in depth the elements that can be included in an archiving strategy.

Niels Brügger, PhD, is an associate professor of Media Studies at the Institute of Information and Media Studies, University of Aarhus, and co-founder of the Centre for Internet Research.

ISBN: 87-990507-0-6

The Centre for Internet Research
Institute of Information and Media Studies
Helsingforsgade 14 · DK-8200 Aarhus N
Tel. + 45 8942 9202 · Fax + 45 8942 5950
cfi_editors@imv.au.dk · http://cfi.imv.au.dk