JANNE NIELSEN

USING WEB ARCHIVES IN RESEARCH

-

An Introduction

Janne Nielsen
Using web archives in research - an introduction

This publication is an updated version of the Danish publication *Forskerbrug af webarkiver – en kort indføring* by the author and published by NetLab in 2015.

ISBN 978-87-93533-00-4

Published by NetLab
Helsingforsgade 14,
8200 Aarhus N.

www.netlab.dk

Contents

Preface……………………………………………..……………………………………….. 4
1 Introduction 6
1.1 The web is not (always) an archive 6
1.2 The special characteristics of web archiving 7
1.3 Digitised, born-digital and reborn digital material 9
2 What is web archiving? 11
2.1 Main types of web archiving 11
2.2 Methods of archiving large amounts of web material 12
2.2.1 Web archiving via web crawlers 12
2.2.2 Harvesting via API 19
2.2.3 Delivery from producers 20
2.3 Frequently-used strategies in harvesting 21
2.4 Interface and search capabilities of web archives 22
2.5 Characteristics of the archived material 23
2.5.1 Reconstructed versions 23
2.5.2 The temporality of the archived material 24
2.5.3 The spatial boundaries of the archived material 26
2.5.4 Other uncertainties 27
3 Existing web collections 28
3.1 Introduction 28
3.2 Netarkivet 28
3.2.1 The collection 28
3.2.2 Strategies 29
3.2.3 Access 31
3.2.4 Search 31
3.2.5 Documentation 34
3.3 The Internet Archive 35
3.3.1 The collection 35
3.3.2 Strategies 36
3.3.3 Access 36
3.3.4 Search possibilities 36
3.3.5 Documentation 37
3.4 Other national web archives 38
3.4.1 Library of Congress Web Archives 38
3.4.2 The UK Web Archive 39
3.4.3 Pandora 39
3.4.4 The Portuguese Web Archive 39
4 The researcher’s own web archive 41
4.1 Introduction 41
4.2 Various methods for creating your own archive 42
4.2.1 Screen capture 42
4.2.2 Screen recording 43
4.2.3 Link crawling (HTTrack) 45
4.3 On-demand archiving 45
4.3.1 Netarkivet 45
4.3.2 Archive-It 46
5 Referencing archived web material (best practice) 47
5.1 Material from Netarkivet 47
5.2 Material from the Internet Archive 48
5.3 References to your own archive 49
5.4 Reference rot – another reason to reference the archived web 50
6 Glossary 51
7 Bibliography 53

Preface

This book has been written in connection with the development of NetLab's workshops on web archiving for researchers. These workshops provide the participants with an introduction to working with archived web materials in research, including a description of what web archiving is, the challenges of using archived web materials as an object of research, knowledge of existing web archives, and tools for micro archiving, so that researchers can themselves archive web materials. The purpose of this book is to gather and make available knowledge about the use of web archives for research. It is written in a Danish context and adapted to the needs of Danish researchers but can also be useful for other researchers.

The book serves as the course material for NetLab's workshops, and is distributed as a free PDF to participants. The structure of the book is therefore inspired by the modules of the NetLab workshops, but it can easily be read independently of the workshops. The book will be continuously updated with relevant new research in web archiving, for which reason it will be made available in several different versions in the longer term.

The original idea was for the book to include manuals for the software programs presented at the NetLab workshops; however, this solution was rejected in favour of making the manuals available at the NetLab website: www.netlab.dk. This has been done in order to ensure to the extent possible that the manuals are always accessible and up to date. In 2016/2017 video tutorials will be added to the website – i.e. brief introductory videos which provide an introduction to the topics of the modules. These can be used both in conjunction with and independently of the book.

Part of the purpose of the NetLab workshops on web archiving is to engage in dialogue with other researchers from the humanities and social sciences who are working with archived web materials in various ways, in order to share and exchange experience. If you possess knowledge of relevant research in web archiving that you think might benefit this book, you are welcome to write to me at janne@cc.au.dk. NetLab's researchers are also in continuous dialogue with the curators and developers at the Danish national web archive Netarkivet (Netarchive), so that the experience gained through NetLab's workshops can be applied in the continued work of developing the Danish web archive as well as digital tools for research using archived web.

Glossary: If you have not worked with web archiving before, a number of the terms used in this book will probably be unfamiliar. The terms are usually explained in the text, but a glossary is also provided at the end of the book, which will hopefully be of help in clarifying them. Terms included in the glossary are coloured *orange* the first time they appear in the text.

1 Introduction

Before we begin to examine the various methods and approaches to web archiving, it is essential to start by asking why we archive web materials at all. This chapter therefore opens with a section on the overall purpose of archiving web materials, and what it is about the web as a phenomenon that makes it so important to try to archive it.

The web today is an important cultural resource as the venue for a large amount of social interaction and many cultural productions, and the forum for much of the public debate. The web plays an important role in society and in people's lives – at least in the Western world. It is therefore important that we learn how to use the web for research, both as a research object in itself and as a source of knowledge about other research objects, and in both contemporary and historical research. However, in order to be able to use the web for research, it is often essential to preserve the web materials, to the extent that this can be done (more on this subject later).

One of the things that can be said to characterise the web is its mutability. The web is a dynamic medium that is constantly changing and evolving at high speed, and new materials are continuously being added. Online, we can receive the latest knowledge and follow events as they unfold, for example by visiting news sites that are constantly being updated throughout the day. The web is also an unpredictable medium. The pace of change is rapid, and it can be hard to predict what types of new online content and services will become the next hit on the web.

1.1 The web is not (always) an archive

Another aspect of the web that is much discussed at the moment is its ability to function as an archive. In 2014, the interpretation of the so-called "right to be forgotten" was the subject of a great deal of debate.[1] In relation to the web, the right to be forgotten has been interpreted as the right of an individual to have sensitive personal data deleted or made less accessible if the material could potentially be stigmatising, and if it is no longer considered relevant. Data can be made less accessible by search engines such as Google removing them from their indices, so that they no longer appear in searches. Discussions and phrasings like these support the notion that the web is a giant archive, in which everything is stored. There is no doubt that there is content on the web that will be stored for many years, and in this respect the web can indeed function as a kind of archive. However, according to Dougherty and Schneider (Dougherty & Schneider, 2011 p. 253) the tendency to view the web as an archive can lead to a lack of understanding concerning the web's dynamic nature. The web is not just an archive, as objects on the web are to a large extent ”continually overwritten,

[1] See for example http://www.theguardian.com/technology/right-to-be-forgotten.

reproduced, reframed, edited, excerpted, and deleted" (Dougherty & Schneider, 2011 p. 253). The web is thus simultaneously characterised by durability and volatility, or as Schneider and Foot describe it, the web is "a unique mixture of the ephemeral and the permanent" (Schneider & Foot, 2004 p. 115). In order to understand the web, therefore, it is important to remember that objects on the web can disappear almost as fast as new ones are added:

> ”The average life span of a Web page is only 44 days, and 44 percent of the Web sites found in 1998 could not be found in 1999. […] As ubiquitous as the Web seems to be, it is also ephemeral, and much of today's Web will have disappeared by tomorrow.” (Lyman, 2002 p. 38)
>
> ”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.” (Brügger, 2005 p. 15)

The figures may be different today, but it is in any case still an important point that material disappears from the web at great speed, as several studies have demonstrated:

> “…several studies found that within a given week 35-40% of web pages changed their content…” (Dougherty et al., 2010 p. 8)

At a presentation given by Brewster Kahle, the founder of the Internet Archive, at an event at the Ford Foundation on 11 February 2015, he stated:

> ”We now know that Web pages only last about 100 days on average before they change or disappear.” (Kahle, 2015)

Because the Web is dynamic, volatile and unpredictable, it is necessary to archive web material if we are to preserve web content and use web material as a research object (Brügger, 2011 p. 24; Dougherty & Schneider, 2011 p. 260; Masanes, 2006 p. 1; Schneider, Foot, & Wouters, 2010 p. 208). The same applies if we wish to ensure that the digital part of our culture that takes place on the web will be accessible – at least to some extent – to posterity.

There are, therefore, many good reasons to archive web materials:

- To maintain our digital cultural heritage
- To stabilize and preserve web materials as a research object
- To be able to document and illustrate claims based on analyses of web materials (whether the web itself is the research object or a source of knowledge about other research objects).

1.2 The special characteristics of web archiving

To archive web materials differs in many ways from archiving other types of media content. If you wish to preserve a book for posterity, you can put it on a shelf in a library, and decide that it should stay there for the future. When you want to use the book again, you can take it off the shelf. It is the same book that has been there all the time, and, apart from some possible wear and tear, it has not changed since it first arrived from the printer. It will also be identical to other copies of the same book that may be on the shelves of other libraries. The same applies if you archive newspapers or magazines. A little more work is involved if you wish to archive radio or television broadcasts, because unlike a book, which is a storage medium in itself, radio and TV content is stored on devices that do not automatically accompany the broadcast. Various storage media can be used, and the choice of medium has a bearing on whether and to what extent the object (i.e. the content you are trying to archive) is affected by the archiving process (for example whether the audio or video quality deteriorates), and how we can subsequently access the content. If, for example, it requires a special machine or a specific format to play the content, then this lays down clear limits for its use, also over time (e.g. obsolete formats, playback machines that are no longer manufactured, etc.).

The way in which you choose to store and make material available thus has a significant impact on what you can subsequently do with the material. The purpose, strategies and technology of an archive affect what is archived and the manner in which it can be accessed, and in this way influence the possibility of constructing a research object on the basis of the material in the archive. This becomes even more evident when we examine web materials, because web archiving is much more complicated than the above-mentioned forms of archiving, as the archiving affects the material much more, and what we end up with in the archive can therefore rarely (if ever) be said to be the same as what we were attempting to archive. We can easily find examples of how archiving, for example, the same *web domain* or *website* in different archives will result in different versions being stored. There are many different methods of archiving, which will affect the object in different ways, with different results.[2] Even if two archives use the same method and archive the material at exactly the same time, different choices can be made during the process in relation to specific settings in the software which can result in versions that are not entirely uniform. The interface, and the functionalities, that are subsequently used in the archive to display the archived versions can also result in significant differences in how the archive’s users can access and process the material. The archiving consequently affects the object itself, and thus potentially also its status as a research object, and since the material cannot usually be viewed in its original context, it is important to understand the significance of this influence. I will return to this point in the section on the characteristics of archived materials (2.5).

[2] See for example Laursen, Brügger and Sandvik (2013) for a description of how four different methods of archiving Facebook data affects what can subsequently be done with the material.

These factors are also very important to keep in mind if researchers wish to archive web material themselves (see 4.1). Different views of what the web actually is will also have an impact on what is seen as being relevant to the archive, and how. Such views are dependent on what it is that we are looking for as researchers (the knowledge interest), and the methods we wish to use to obtain knowledge of this. One could for example focus on aspects of content (what it says, what it is about), or on layout elements, functionalities, linking structures between websites, or interaction and use. There are differences in the types of data you will need if you wish to undertake an analysis of the layout or of the content of a website such as www.dr.dk, or if you wish to undertake a network analysis of the network of sites that link to www.dr.dk, or that www.dr.dk itself links to. So here, too, archiving and the choices made in this connection may be of significance for the research object you end up with.

1.3 Digitised, born-digital and reborn digital material

Web archiving is basically the archiving of web material (I will give more specific definitions later), and web material is – like many other media today – digital in form. But web material that is archived is digital in a particular way, and this plays an important role in our understanding of the archived web. We can basically distinguish between three types of digital material (Brügger, 2012 p. 104):

- Digitised material which has previously been analogue material, and which has been converted into digital files (consisting of binary code).
- Born-digital material that was digitally produced from the start, and thus has no analogue original.
- Reborn-digital material, which is created when digital material (digitised or born-digital) is collected and preserved in a process in which the material is altered (Brügger, 2012 p. 104).

The material found in web archives could be described as reborn-digital, because what is stored in the web archive will never be quite the same as the material found on *the live web* (i.e. what is online now). This will be explained further in the section on the characteristics of the archived material (2.5), but it is important to emphasise here from the start. There is thus a significant difference between reborn-digital material and other types of digital material, and this has a crucial impact on how we can understand and work with archived web materials.

When we work with web materials it is also appropriate to define how we talk about them, and this book, using concepts borrowed from Brügger (2009), distinguishes between three levels: web element, web page and website. A web element is the smallest meaningful unit on a web page, such as a defined piece of text, a picture, a graphic element, a video or the like.[3] A web page is what is viewable by the user in a browser window (including what can be seen by scrolling down), while a website consists of a number of contiguous pages (ibid.). Brügger points out that ”the website is based on a semantic, formal and physically performative coherence” (Brügger, 2009 p. 123). Connections are thus created not just through form and content-related (semantic) elements, but also through the structures (links) that connect the elements via actions performed by the user (for example when by clicking on a link you move between the various parts of a website). As users, we usually have a clear understanding of what a website is, and we can distinguish between, for example, TV2’s website and DR’s website (Danish broadcasters), which are located on separate domains, and which each have their own layout-related characteristics and other distinctive aspects that ensure their individual coherence and distinguish them from each other. A website may also consist of several sub-websites like http://nyhederne.tv2.dk/, http://sporten.tv2.dk, http://vejret.tv2.dk, etc.

It must be emphasised that these definitions of web page and website are to some extent simplifications that we use in order to make it possible to separate objects analytically and talk about them. The boundaries between a website and a web page are actually much more complicated – and the complexity increases when we talk about archived versions of web pages and websites, because the interlinked structures of the web make it hard to accurately delineate these units. At the same time, one can discuss the extent to which it is necessary to try to capture the entire context in order to preserve that which we wish to preserve (see 2.5). With this in mind, we will now turn to the question of what web archiving actually is.

[3] In some cases the term web object is also used, since it is the term used by Netarkivet to denote the smallest unit of the web archive.

2 What is web archiving?

2.1 Main types of web archiving

The International Internet Preservation Consortium (IIPC), a member organisation for web archives that works to promote international co-operation on web archiving and improve its tools and standards, describes web archiving as:

> ”… the process of gathering up data that has been published on the World Wide Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research.” (http://netpreserve.org/about-us)

The need to make material accessible forms an important part of this perspective, which is natural considering that the members of the International Internet Preservation Consortium are primarily archiving institutions preserving cultural heritage. However, web archiving can also be performed by individual researchers or research groups. Brügger operates with a broader definition of web archiving:

> ”Any form of deliberate and purposive preserving of web material.” (Brügger, 2011 p. 25)

Brügger also distinguishes between macro archiving and micro archiving (Brügger, 2005; 2011). Macro archiving, understood as archiving on a large scale, is typically performed by institutions whose aim and task is to archive cultural heritage, and who therefore possess the technology and the expertise to archive large amounts of material independently of individual research projects. Here, the aim is often to preserve as much of the cultural heritage as possible, without having a specific research purpose or a specific need in mind – a form of ”preservation for its own sake”, as Thomas, Meyer, Dougherty, Van den Heuvel, Madsen and Wyatt (Thomas et al., 2010 p. 7) call it. At the same time, this means that the archived material will in this case often be so varied, and cover so many topics, that it can be used for many different types of research projects – provided, of course, that it is archived in a form that proves useful to the researchers who subsequently access it.

Micro archiving, on the other hand, takes place on a smaller scale, and is often highly focused, as in this case an individual researcher or research group archives precisely the material they assume they will need in relation to a specific object of research. It is therefore often based on a here-and-now need to stabilise a given object of research and preserve it in such a way that it can be utilised for the specific needs of the researcher in relation to research questions, methods, etc. (Brügger, 2005 p. 10; 2011 p. 25). Unlike macro archiving, micro archiving can be carried out by researchers who have no particular experience with web archiving or especially advanced software. Brügger points out that web archiving must in any case be understood as a deliberate and purposeful action, in which choices are made on the basis of an intention to archive, together with reflections on why this material should be archived (Brügger, 2011 p. 25).

To these two categories can be added initiatives in which one or more institutions undertake archiving on a scale that can neither be described as macro or micro archiving, but as something in between, which we might call thematic or selective archiving.[4] Here, the goal is often to build collections on specific topics or areas on the basis of some ideas about what it may be particularly relevant to archive, but without wishing to archive everything, as in macro archiving, and without necessarily having a specific purpose, as in micro archiving.

All forms of archiving have their advantages and disadvantages, and are therefore best suited to certain types of purposes. It is important to be aware of how the various kinds of archiving affect the end product and what you can do with it. Another crucial factor is which file types are included in the web material to be archived. There are differences in the methods required to save static and dynamic content, respectively, and highly dynamic content (such as YouTube videos, social media, etc.) require different or special techniques. In the sections that review the various methods, you can also find descriptions of their various advantages and disadvantages.

There are not necessarily agreed definitions of the different approaches to web archiving. Thomas et al., for example, write that one traditionally distinguishes between selective harvests and domain harvests, with the former mainly carried out by individuals or groups, and the latter typically undertaken by archives and libraries (Thomas et al., 2010 s. 9). The Danish national web archive, Netarkivet, however, makes use of both selective harvests and domain harvests (known as bulk or snapshot harvesting, or broad crawls – cf. later sections). There can thus be large or small differences in what is meant by these terms in practice.

2.2 Methods of archiving large amounts of web material

There are several different ways to archive web material. Some methods are particularly suitable for micro archiving, while others are suitable for macro archiving (or thematic archiving). In the following, we will focus on the methods of harvesting used in macro archiving. Micro archiving methods will be discussed in the section on the researcher's own archive.

2.2.1 Web archiving via web crawlers

One of the most popular methods of archiving web material, when you are looking for more than just individual pages and wish to archive the hyperlink structures as well, is what is known as *web crawling* or *link crawling*. This is the method most

[4] Several of the institutions that carry out macro archiving also make use of selective and/or thematic filing as part of their strategy to collect as much as possible of the cultural heritage (see 3.2.2 and 3.4).

commonly used by web archives (Thomas et al., 2010 p. 10), which harvest on a large scale, but it can also be used for micro archiving if a researcher wishes to archive a single website including link structures.

The method can in general be described as ”recursively following embedded hyperlinks to some depth” (Thomas et al., 2010 p. 10), or in a slightly more detailed definition as ”the process of collecting web material and loading it into a fully browsable web archive, with working links, media etc.” (Laursen, Brügger & Sandvik, 2013). A program called a web crawler or spider is made to systematically move around the web by following links, and on the way collect, analyse and download information from web servers. Such crawlers are a kind of *web bot*, i.e. a software application that performs automated tasks on the web, and which can be used for many types of tasks, including indexing, updating and archiving. When we talk about web archiving, a crawler is often described as a ‘harvester’. All web archives that are members of the International Internet Preservation Consortium (IIPC), including Netarkivet and the Internet Archive, use Heritrix, which is a flexible and scalable harvester (Library of Congress, n.d.). It was developed by technicians from the Internet Archive in co-operation with technicians from various institutions affiliated with the International Internet Preservation Consortium (ibid.), and is available as open source software. The software is continually being developed, and individual archives have the option of tailoring Heritrix to their specific needs.

The harvester is assigned a list of *domains* or *URL*s (web addresses) to archive. The harvester begins at a URL, and archives as many web objects as possible by downloading files from the server(s) in question to the archive servers. All of the internal and external links (URLs) are also harvested, after which the harvester moves on to the linked pages, archives them and harvests their links, then moves on to these URLs, and so on.

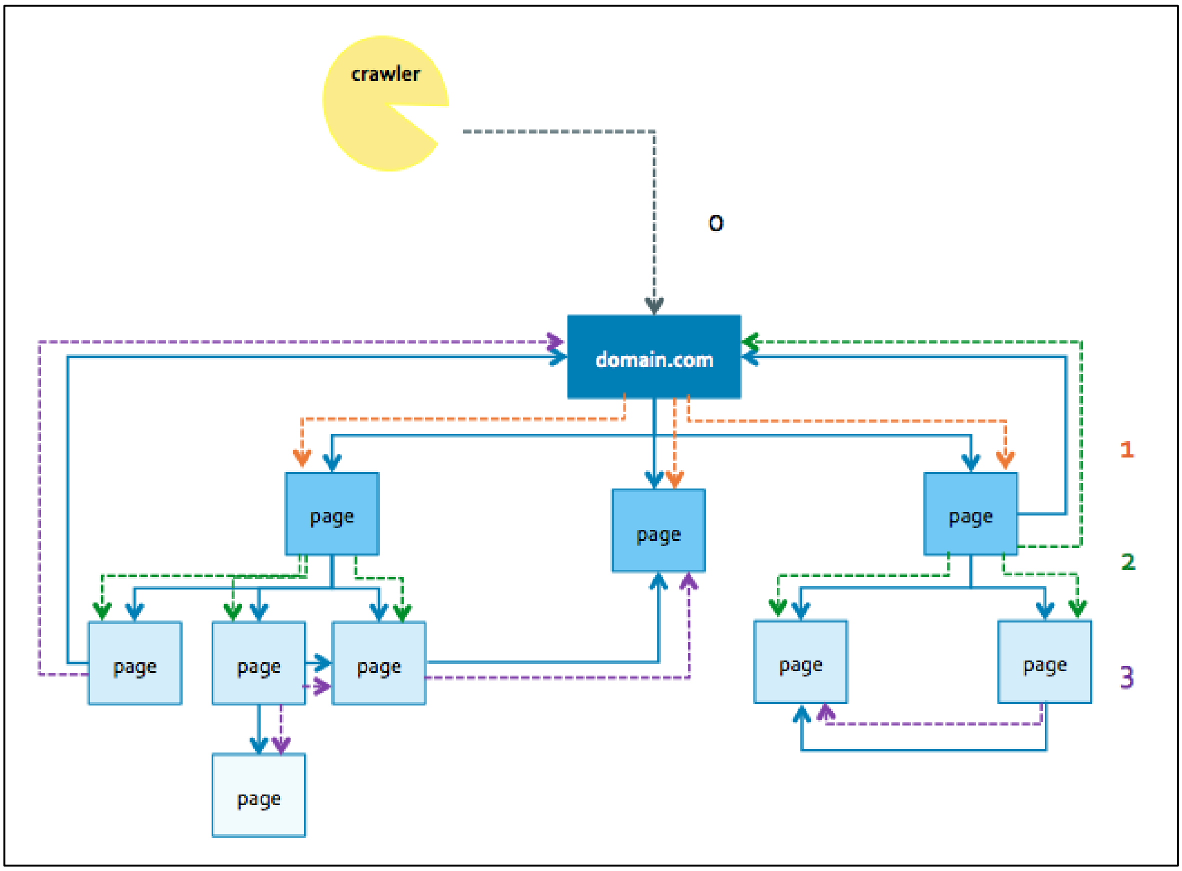


Figure 1: The harvesting process. The blue lines represent links, while the dotted lines show the harvester's movement (at various levels).

The harvester also archives metadata along the way, which is stored in the so-called 'crawl log'. This logs both the resources collected and the progress of the harvesting. The person who configures the software must tell the harvester in advance how to handle many different parameters, such as how many levels should be harvested (i.e. how many times the harvester should follow links further from the original URL, cf. the numbers and colours in Figure 1), whether there is to be an upper limit on how many objects or bytes are to be harvested in each domain, whether certain file types are to be harvested or not, etc. The way in which these parameters are set is very important for the harvesting process, both in relation to the length of time it takes (and how complicated it is), and in relation to the archived material that you end up with.

Different methods offer different possibilities and challenges. Harvesting with link crawlers has the great advantage of ensuring that entire web pages are collected, that the relations between web objects (both the interconnection of objects on each page, and the hyperlinks that connect across URLs) are preserved, and that the archived material can subsequently be displayed in an interface in which it looks and behaves like the live web (with certain important exceptions, which I will describe below). As the harvesting can be partially automated, it also has the distinct advantage of being able to collect large amounts of material.

Although harvesting by link crawling has great advantages, the method also has significant challenges. First of all, the harvester may encounter links that do not work and content that no longer exists – but the same problem can also be encountered with other methods of archiving. A bigger problem is that several types of objects cannot be collected and archived by the harvester, and that various things can stop the harvester (as well as other crawlers) on its way around the web.

Firstly, there is a very large part of the web to which the harvester cannot gain access. Some estimates suggest that up to 90% of all content on the Internet is inaccessible (Dougherty et al., 2010 p. 8), because it is stored on the part of the net not indexed by search engines – the so-called deep web or hidden web.[5] None of this content can normally be harvested. I will not go into greater detail about the deep net, but just mention that it consists amongst other things of a huge volume of databases, FTP servers, private web pages and other password-protected content, content intentionally hidden from crawlers, pages that are not linked to other pages, and dynamic content generated on the basis of queries.

Dynamic content is in general a major challenge for harvesters, and even when some of it can actually be archived, objects with various types of dynamic content may create problems. JavaScripts that download content from a server, for example, represent a major challenge. The script can be archived, but when the archived script is 'replayed' in the archive interface, the script often fails because the content from the original server (which is a remote server and not one of the archive servers) cannot be retrieved. Alternatively, it may happen that the script can indeed retrieve the content, but the content is no longer the same as it was when the object was archived. An example might be a script that retrieves the day’s weather forecast, exchange rates, or the like, and it might for example mean that an element which has been retrieved today can suddenly appear as part of a web page that is several years old – without the user necessarily being aware of this. In general, objects that create content as a result of a look-up in a database (which is located on the original server) represent a major challenge. Basically, we can say that if a functionality on the web contacts the original server and requires some form of action by the server, it will fail in one way or another in the archived version.

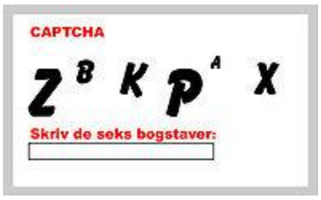
Content based on Flash pages and interactive social media comprises dynamic content that cannot be archived using harvesters (Schostag & Fønss-Jørgensen, 2012 p. 120). The same applies to video, audio and streamed elements, for which reason Netarkivet regularly launches special initiatives to harvest videos online (Tue Larsen, Director of Netarkivet, The Royal Library, personal correspondence, 03.09.2014). The archiving of these types of content is complicated by the fact that

[5] Unlike the so-called ‘surface web’, which is indexed.

websites like Facebook and YouTube are continually changing their formats and options (Thomas et al., 2010 p. 10), as a result of which archiving institutions must continually develop new methods to attempt to collect this content.

Other types of functionalities and dynamic content, namely the kind generated on the basis of the user's context, and which possesses a form of complexity that the user does not notice in everyday use (Day, 2006 p. 193), are not archived either. Today, a large part of what we see online as users is dependent on our software (e.g. our browser and possible plug-ins) and on our past behaviour, as cookies store our behaviour and data for subsequent use in targeting content, such as banner ads, at the individual user. Content like cookies, banner ads, comment sections, sharing functions, plug-ins, etc., are not usually seen as being part of the primary content on the web, but as Rogers (2013) points out, they represent an essential part of websites and of the functionality of the web that is rarely archived.

Websites with various kinds of access restrictions, such as password protection or requirements for IP authentication, can quickly put a stop to the activities of harvesters, as can websites that require "nontrivial interaction" (Masanes, 2005), i.e. where a user must enter certain things or perform actions that a crawler cannot, such as captcha codes, which are the small forms that some websites require you to fill in to prove that you are a human.[6]


CAPTCHA

Skriv de seks bogstaver:

Picture from: http://da.wikipedia.org/wiki/CAPTCHA

In addition to items that are difficult or impossible to archive, items on the web can impede the harvester's journey around the web, or possible disable the harvester. Some types of content can act as a kind of trap for the harvester; the so-called *bot traps* that generate links, thereby creating an endless loop of requests that causes the harvester to move in circles and/or crash. An example is an online calendar, which can have an infinite number of links. Another major barrier to at least some harvesters is the so-called *robots.txt exclusion*. Robots.txt is a de facto standard that can be added at the root of a domain, instructing automated systems not to crawl a site or parts of it, so as to prevent the crawler's requests overloading the site's servers unnecessarily. It can be helpful, for example at a site where the structure creates problems (cf. bot traps), or where the content is inappropriate for most crawlers, such as search engine harvesters – but it can also be a way to try to prevent content being crawled and archived, for various reasons. Ordinary web etiquette is to follow the directions given – also because, as mentioned above, it may be beneficial to the crawler to do so.

[6] Captcha stands for Completely Automated Public Turing-test to tell Computers and Humans Apart.

The Internet Archive respects robots.txt, also retroactively, in the sense that the Archive not only does not harvest a domain that has added robots.txt, but also blocks the display of any previously-archived versions (created prior to the addition of the robots. txt file) (see https://archive.org/about/faqs.php#14).

In Denmark, however, the legal requirement for published works to be deposited applies to all publicly-available content on the Internet, for which reason Netarkivet's harvesters do not respect robots.txt:

> "It is not acceptable that the harvester skips any material that is subject to the deposit obligation, so the harvester used must not, therefore, follow any instructions to do so. In accordance with this, the legal deposit institution will ignore current norms to prevent harvesting, such as robots.txt."
> (Pligtaflevering.dk, n.d.)

Netarkivet has made previous attempts to respect robots.txt to test whether the data that was then not archived was relevant. It was found that when the harvesters followed the instructions in robots.txt it resulted in a significant reduction in the volume of the archived data, the content of which could largely be shown to be irrelevant (Netarkivet newsletter, March 2011[7]). However, it also turned out that some relevant material was omitted from the harvesting, for which reason, in order to meet the requirements of the Act on Legal Deposit of Published Material, it was decided that robots.txt would not be respected.

As the above has hopefully illustrated, many challenges are associated with harvesting using crawlers, and new challenges are continually arising. In web archives, harvesting is carried out by technicians and curators who continuously monitor the crawler and work to deal with the challenges it meets. For these heritage preservation institutions harvesting is a race against time, not only because websites are changing, but also because the constant development of the internet in general makes it almost impossible to predict what the next big challenges will be for web archiving. In the archives, curators and developers must constantly keep themselves up to date on new forms of content and services, and continually work to develop technology and software for the collection of new types of content (Schostag & Fønss-Jørgensen, 2012 p. 119).

In conclusion, we can sum up the pros and cons of harvesting via web crawling. In doing this, I draw on a very important contribution in describing different web archiving methods and their advantages and disadvantages: In the paper "Methods of collecting facebook material and their effects on later analyses" (Laursen, Brügger & Sandvik, 2013) Ditte Laursen, Niels Brügger and Kjetil Sandvik discuss the process of archiving web and how different methods influence the material and what you can do with it. The methods are discussed in relation to

[7] http://netarkivet.dk/wp-content/uploads/Nyhedsbrev_Netarkivet_2011_Marts.pdf

the process of archiving Facebook material but the advantages and disadvantages will be similar for many other types of web material and the overview provided by the paper is, therefore, very useful and relevant beyond the specific scope of the paper. In the following chapters, I draw on the results of Laursen, Brügger & Sandvik (2013) whenever I mention a list of pros and cons, as well as in the broader description of the advantages and disadvantages. My focus, though, is not only on Facebook material but on web material more generally, so I attempt to make the points more generally applicable, and the following lists of pros and cons can, thus, be used when deciding on methods for other types of material as well. In some cases, I elaborate on the advantages and disadvantages described by Laursen, Brügger & Sandvik (2013), and in some cases I add pros or cons, but the results from "Methods of collecting facebook material and their effects on later analyses" (ibid.) has informed all pros and cons list in this publication.

Pros of web/link crawling:

- The full website can be archived (assuming that the harvester is set up to follow all links throughout the website), and what is archived usually resembles what was online to a large extent.
- The full length of the individual page is preserved.
- The link structure is preserved, and thus the interrelations between web elements and pages, as well as other websites.
- The archived material looks and behaves like the live web (with some important exceptions). The archived version is displayed in a browser, and it is clickable so you can move around by following the links, just like on the live web (except for the temporal issues).
- The html is archived, which means that the archived versions are machine-readable, which provides good possibilities for searching and sorting, and enables links to be clickable.
- It can be performed automatically (in part, as there is a need for ongoing monitoring to evaluate the collected material and deal with any technical difficulties that arise)
- Access to metadata (crawl logs)
- Can be used for big data analyses (e.g. content analysis, network analysis, etc.)

Cons of web/link crawling:

- The archived version does not necessarily look exactly what was online as some objects cannot be archived, such as videos and streamed content, as well as applications that use Flash, JavaScript, etc.
- Content that requires user interaction cannot be archived.
- Difficult to spatially delimit
- Temporal inconsistencies
- Risk of the harvester getting caught in 'bot traps' (requires some monitoring)

(Laursen, Brügger & Sandvik, 2013).

2.2.2 Harvesting via API

Social media represent a major challenge for web archives. They play a very large part in the web behaviour of most Western internet users, and much of the public debate now takes place in these fora. They must therefore be regarded as an important part of our digital heritage, but they present significant challenges to the harvester, partly because much of the content lies behind password protection. Some of the contents of Facebook, namely the publicly-available pages, can at the present time be harvested by Heritrix. The caveat "at the present time" is important, because Facebook frequently changes its software, which increases the challenge of collecting data.

As much of the social interaction on Facebook takes place in closed groups, on people's private 'walls' or in feeds, researchers who wish to investigate aspects of the interaction on Facebook need to obtain access to the closed fora. With support from NetLab, a software application called *Digital Footprints* has been developed which can harvest data from Facebook using the Facebook API (Application Programming Interface). An API is a software interface that makes it possible to extract data from one system and make it available in another system. In August 2006 a beta version of the Facebook Development Platform was introduced,[8] and from May 2007 onwards, it became possible for software developers to develop applications on Facebook's platform that could be used in conjunction with Facebook (Brügger, 2013a p. 33). The API makes it possible to collect data from a user's profile, which is then used to customise the application to the user, and to publish information about an app on the user's profile, such as news feeds showing that one of your friends is playing a certain game or has run a certain number of kilometres. This data is recorded by an app and can then be shared with the user's Facebook friends (depending on the user's settings, such as privacy settings). Not all information, however, is available through the API:

> "The Facebook API suite makes available a number of different (but not all) aspects of the Facebook collection of objects, albeit bound about with various security constraints." (Thomas et al., 2010 p. 19)

However, some data may be obtained, and because the API can be used for this, it can also be used in a research context. The above-mentioned *Digital Footprints* program can, once the researcher has received permission from both a specific user and the Data Protection Agency, download the user's data (profile information, etc.) and store this in a database. The researcher can access this data in the Digital Footprints user interface, which offers several different views, for instance the news feeds or profile feeds from participants, or the feeds from pages or

[8] https://www.facebook.com/notes/facebook/facebook-development-platform-launches/2207512130

groups. The timestamps of all the posts are available so it is possible to follow the activities over time. Digital Footprints also offers some Facebook statistics so the researcher can see statistics of the types of post and comments. It is also possible to search in the data and generate statistical analysis on this basis.

It is important to note that Facebook regularly changes its data access policy, which can affect what it is possible to access for researchers through the API. Digital Footprints can also be used to retrieve Twitter and Instagram data. Only university researchers (including PhD students and postdocs) can use Digital Footprints, and access is only granted when applied for in relation to a specific research project (http://digitalfootprints.dk/project/application).

Other ways to obtain data from social media exist; Twitter, for example, provides several possibilities for accessing data, which can be seen in the developer section of their website (https://dev.twitter.com/streaming/overview).

Laursen, Brügger and Sandvik (2013) provide a good summary of the pros and cons of collecting web data via the API. The summary is based on an attempt to collect Facebook data using Digital Footprints.

Pros of archiving via API:

- The entire content of the page is preserved.
- The development over time is preserved.
- The data preserved are machine-readable, which means that they are searchable, clickable, and sortable.
- The API provides access to data that would otherwise be hidden.
- The data can be handled as big data.

Cons of archiving via API:

- The original visual appearance of the elements is not preserved.
- Videos cannot be played (instead, links are provided to the live web).
- Streamed content is not collected.
- The content that is linked to is not collected (links are instead provided to the live web).
- Proprietary format.

(Laursen, Brügger & Sandvik, 2103)

2.2.3 Delivery from producers

One could also imagine a solution in which libraries could have an agreement with owners of websites to provide material, in the same way as with other forms of material covered by legal deposit. Such an agreement could for example be utilised if web archives wished to archive material from web radio stations that are only available online, or where streamed content presents a challenge for the harvester, or in relation to e-books on the web, etc. The Internet Archive, which is not a legal deposit institution, has received some material supplied from other sources in this manner, including both web material and other types of digital material.

2.3 Frequently-used strategies in harvesting

As can be seen from the above description of the challenges of harvesting, it is not possible to preserve everything, both because harvesting is done in a ‘friendly’ manner so as not to overload the servers from which the data is harvested, and because not everything can be harvested using the various methods. Given that it is not possible to harvest everything, then, it is necessary to choose some strategies for what is archived, and how. This applies to all types of web archiving. A single strategy is often not enough if you wish to ensure both breadth and depth in the archiving, which is why it is common practice to combine different methods – at least in the web archives operated by libraries and similar heritage institutions. This section will briefly review the most common strategies for archiving web material when it is necessary to archive large amounts of material through harvesting (using crawlers). Some of these strategies will be elaborated in the section on Netarkivet (3.2), which utilises several of these strategies. Other strategies will be applicable in the case of micro archiving, as shown in chapter 4 on the researcher’s own web archive.

Broad crawl (or bulk/snapshot harvesting): A form of broad harvesting that attempts to harvest more or less everything, or at least as much as possible. The word ‘snapshot’ is in some ways misleading, because it suggests that the material is recorded at a given moment in time, which is not the case. First of all, it may not be possible to save a complete snapshot of the network, and secondly, this type of harvesting usually takes a long time – up to several months (see 3.2.2 on Netarkivet’s strategies). Masanes (2002) therefore points out that it is ironic to speak of a ‘snapshot’ in connection with this type of harvesting operation.

National or regional domains: Harvesting that focuses on selected top-level domains, such as .dk or .uk. The data collection is usually done through broad crawls.

Selective harvesting: Harvesting of specific domains. Here one might for instance focus on harvesting in even more depth than is usually done in broad crawls.

Event harvesting: In this type of harvesting, an attempt is made to harvest all websites that are relevant in relation to a particular event. Some event harvesting may be planned in advance, e.g. harvesting all websites in relation to a general election or the Olympics, while other events, such as natural disasters or terrorist attacks, are unpredictable, in which case harvesting can only be initiated when we become aware that the event is something that ought to be preserved for posterity.

Thematic harvesting: Similar to selective harvesting, but centred on a particular theme or subject area, which is considered particularly important to preserve. This can be similar to event harvesting, but is not necessarily restricted to the time limit of an event.

2.4 Interface and search capabilities of web archives

The objects and files obtained from harvesting, and which are preserved in the archive, may be presented in many different ways. There will usually be differences in how the technical staff at the archive can access data and metadata, and the options available to users to obtain access to the material. One way to present archived versions of websites that is used by many of the web archiving institutions is the *Wayback Machine*.[9] The Wayback Machine is an open source software application, originally developed for the Internet Archive by Alexa programmers (Kimpton & Dubois, 2006),[10] and which is being developed on an ongoing basis (currently under the name 'Open Wayback'). The Wayback Machine is intended to 'replay' material in web archives, which is to say that the software downloads and assembles the archived objects that make up a web page, and replays them (see also 2.5 on the archived web as a reconstruction). The software rewrites all the links (Taylor, 2012) so that they link to and from archived resources. In this way, the software can 'surf' the archive in the same way as is familiar to us from the live web. You can also 'time travel' in the archive, for example by jumping between different versions of the same web page that have been archived at different times. The search function of the Wayback Machine is based on URL lookups, i.e. access to the archive can only be achieved by searching for a specific URL. (For more about the functionality of the Wayback Machine, see 3.2.4 on the search options in Netarkivet.)

In general, most current interfaces for web archives are based on some kind of URL look-up, possibly combined with simple searches, for example relating to a particular period of time (Thomas et al., 2010 p. 23). However, some archives have permitted other types of searches for some time now, and ongoing efforts are being made to enhance access to the archives. One obvious possibility is a free text search, which provides quite different possibilities for finding material, but can also potentially create new challenges such as data overload (also for users). The search possibilities will often be linked to the size of the archive: the larger the

[9] According to Taylor (2012) the Wayback Machine is the most widely-used software "used to 'replay' the contents of ISO-standard Web ARChive (WARC) file containers" (ibid.), which is the format used by the Heritrix harvester.

[10] Alexa Internet worked closely with the Internet Archive, especially in the first years of the archive's history (Kimpton & Dubois, 2006).

archive, and the faster it grows, the harder it will be to index, which is the prerequisite for free text search.[11] In the same way, the harvesting strategies used can also influence subsequent search options, as we see in some archives, where you can browse various topics, perhaps arranged in alphabetical order. This makes good sense in collections gathered by thematic harvesting, in which the archived versions of websites are already defined as belonging to a theme or topic.

A web archive will also usually contain various kinds of metadata that are important in the development of new search possibilities. You could, for example, work using several types of data from WARC files, crawl logs, indices and the like in order to search in different ways in the archives. It also depends on what kind of data you are interested in accessing, and what you wish to do with it.

2.5 Characteristics of the archived material

Many challenges are associated with web archiving: technical, economic, legal, etc. However, some of the greatest challenges, seen from the researcher's perspective, come down to two factors: firstly, that it is impossible to save everything, and that the choices made are significant for the object:

> ”Web archiving is often a matter of choices, as perfect and complete archiving is unreachable.” (Masanes, 2005 p. 77)

And secondly, that the object you are attempting to preserve when you create a web archive will in most cases be distorted by the actual archiving process.[12] It could be argued that it is impossible to completely preserve web materials. However, we must still try, because the alternative is that we cannot use the web for research, to put it bluntly. You might expect that, which is in the archive to be a copy of what was on the live web, but this is rarely – if ever – the case. The researcher cannot therefore know whether the item found in the archive looked exactly the same when it existed on the live web (Brügger, 2005), or whether the various parts of the archived version of a website would ever have existed simultaneously. In order to create the best conditions for research using the archived web, we must therefore be aware of what characterises the archived web, and what precautions we should take.

2.5.1 Reconstructed versions

As mentioned to begin with, archived web can be described as 'reborn-digitised' material, due to the changes that follows from the archiving process itself:

[11] There may also be some associated legal issues, as free text search allows you to combine data in other ways than a URL lookup, and thus presents greater challenges in relation to, for example, personal data protection (for the archives and countries who care about this!).

[12] Parts of this chapter build upon the chapter *Arkiveret web som forskningsobjekt* (The archived web as the object of research) in my PhD dissertation (Nielsen, 2014).

> ”…the process of archiving creates the archived web on the basis of what was once online: the born-digital web material is reborn in the archive.” (Brügger, 2012 p. 108)

When you enter a web archive and view an archived version of a website (on the basis of a specific URL and a selected time code), what you see is a reconstruction, *not* a copy of the site. The reconstruction is created in two steps. Firstly in the archiving process (harvesting), during which many choices are made about strategies and tools (URLs, depth, width, file types, time, the handling of objects that cannot readily be harvested, etc.)[13] which will affect how the material in the archive may come to look, and secondly, the interface used to make the collection (i.e. the material in the archive) accessible. This could be an interface that simulates a web server, which supplies content that appears in a browser, as the Wayback Machine does. When the user sends an inquiry to the Wayback Machine by choosing one of the archived versions from the results (calendar view) that are obtained by searching on a URL (see 2.4 and 3.2.4), the Wayback Machine creates a version by assembling the relevant objects that have previously been harvested and archived, and thereby stabilised (Brügger, 2010 p. 7; Schneider et al., 2010 p. 213). The manner in which the Wayback Machine software is structured affects the object, and we can therefore call this version a reconstruction. This does not, however, mean that it is a completely random collection of objects, because the composition is based on the HTML code (the language in which the website’s structure is described), which was harvested.

2.5.2 The temporality of the archived material

Researchers who work with web archiving often point out that the web changes very rapidly and without necessarily following any particular logic, and that you cannot see what was there before, because updates often overwrite previous versions. Brügger speaks of “the dynamic of updating” (Brügger, 2005; 2008; 2009; 2010; 2011), and Schneider & Foot (2004) and Masanes (2005; 2006) point out that we cannot know when something has been updated, and it is therefore difficult to chronologically date web material.

> ”What is harvested is both a *point* in time (the time of harvesting) and a *period* of time (the period up to the time of harvesting).” (Brügger, 2008 p. 158)

The material can be assumed to have existed at the time of harvesting, but it may have looked like that for a very long or a very short period of time, depending on the length of time between the updating and the harvesting – and we cannot know how long this has been. If a site is archived very frequently, this may of course help us by allowing us to compare different archived versions, but even with sites that are harvested very often, it will probably not be possible to capture all of the changes and clarify exactly when they came into being, as these sites are typically

[13] Cf. Brügger (2005), Masanes (2005) and Schneider, Foot & Wouters (2010) on web archiving strategies.

characterised by a high update rate (which is why they are harvested so often – see the description of selective harvesting operations in 3.2.2). There may of course be exceptions, such as when objects on the site are explicitly marked with the date and time, as we often see on news sites. In the past, some websites used to indicate somewhere on the page when it had last been updated – see for instance the versions of some of the early pages of the website www.dr.dk as they are archived in the Internet Archive – but very few pages continue to do so today. Now the expectation is rather that, unless otherwise indicated, the information presented on a website is always current.[14]

Another factor that it is crucial to understand when we try to determine the temporality of archived web objects is that the harvesting process itself takes time, and that this influences the archived object. This is particularly important to keep in mind if we move away from a single web page and instead look at an entire website. Here, the various parts of the website may have been harvested at different times. The uncertainty is especially significant with large, highly dynamic sites, as it may be assumed, firstly, that it takes longer to archive than a smaller site, and secondly, that the objects on the site have most likely appeared, been removed or been updated while the harvester has been archiving the site. This means that the version of a website that is found in an archive may be composed of objects that were not all online at the same time. Brügger provides a very good example of this in his description of an archive that he created:

> "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 half an 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)." (Brügger, 2005 p. 22)

It is therefore important that we speak of a version of a website, not only because there may be multiple versions of the same site from the same day, but also because what is found in the archive is always a version, because it can never be a complete copy of what was once online (Brügger, 2008 s. 161; 2009; 2011 pp. 33-34; Brügger & Finnemann, 2013).

Because of the potential asynchronicity between updating and archiving, the reconstruction (the archived version) may contain either more or less than what existed on the live web at any given time (Brügger, 2011 p. 33). It will rarely be possible to find a version that does not have parts missing, but at the same time there may also be objects that are, so to speak, in excess, because the individual

[14] "Om DR Online"_DR Online_29 July 1997 [http://www.dr.dk/omonline.htm]. _Internet Archive_. [http://web.archive.org/web/19970729014603/http://www.dr.dk/omonline.htm]."

objects from which the reconstruction is constructed are not from exactly the same time. Brügger describes this paradox:

> “On the one hand the archive does not look like the internet as it *actually* was in the past (we have lost something), but on the other hand the archive might look like the internet as it *never* was in the past (we get something different).” (Brügger, 2001 p. 6)

The material is therefore incomplete, because some parts are not included, and it is difficult (if not impossible) to know precisely what is missing. But in a sense there can also be too much material, in that the reconstruction may contain more material than what was ever actually on the site, due to the duration of the archiving process, or because multiple versions of the site existed at around the same time (Brügger, 2013b p. 314). This happens, for example, when there are several web pages that link to the same URL (e.g. www.dr.dk), which are then harvested repeatedly.

If, as a researcher, you archive material yourself, it is possible to examine during the process the extent to which the archived versions you collect match what is on the live web, and what you are attempting to archive. If, on the other hand, you utilise material from other web archives, you do not have the same opportunity to check whether the various archived versions are more or less in line with what was on the live web at any given time (Dougherty et al., 2010 p. 23). One way to learn about possible differences could be to compare versions in different archives (see for example Masanes, 2005), but even if these are similar, it still does not eliminate the possibility that they might all lack something that was on the live web but was not included in the archiving. As what we have access to are not identical copies, but potentially inconsistent versions (Brügger, 2009 p. 125), and since it is difficult to clarify any inconsistencies between what was on the web and what is in the archive, it is important for researchers to be cautious when using archived versions of web material as a research object.

2.5.3 The spatial boundaries of the archived material

It may therefore be difficult to temporally delineate a research object when working with web material. Spatial delineation presents a further challenge (Brügger, 2009 p. 128; 2012 p. 109; Brügger & Finnemann, 2013 p. 75). One of the basic characteristics of the web is the hyperlinks that bind websites together. The boundaries of a website can therefore be said to be unclear – for where does a website actually start and end, when we consider the link functionality?

> “The average Web page contains 15 links to other pages or objects and five sourced objects, such as sounds or images. For this reason, the boundaries of the digital object are ambiguous.” (Lyman, 2002)

A book has clear physical boundaries, but a web object has not. This not only creates challenges in relation to the researcher’s description and delineation of the research object, but also in the actual archiving of a website, where, according to Lyman, one should ideally seek to preserve the structures and the experience that characterises the Web:

> "…an archive also must be sure that the document is *translated* in an authentic manner. In this case, authenticity means that the document must both include the context and evoke the experience of the original." (Lyman, 2002 p. 41)

Thomas et al. also point to the experience of the web as something potentially different to and greater than just the content on the web:

> "A distinction can be made between capturing the *content* and capturing *the way that the content is experienced*. An archiving strategy must decide whether to capture only content, or to attempt the much more difficult task of capturing the appearance(s) and behaviour(s) in all their possible varieties." (Thomas et al., 2010 p. 10)

Thus, in the delineation of a web object, many aspects are relevant to take into account at several stages of the process, including when we view the archived material.

2.5.4 Other uncertainties

One last thing that should be mentioned here, and which is important to remember about archived web material, is the possible sources of error that relate to the challenge of archiving dynamic material, as mentioned in the section on web archiving via harvesters (2.2.1). If an archived version of a web page includes a script that downloads dynamic materials, there is a risk that part of the content that you see in the archive is being downloaded from a database on the original external server at the same time as the archived page is being replayed, and that this material, therefore, does not match the temporality of the other objects displayed. If the script for instance retrieves the current weather forecast, then the interface shows the weather at the present time, even though the archived version of the page was harvested three years ago, and all content on the page should therefore be three years old. Another potential source of error that could be reflected in the archived material is seen if all CSS code (cascading style sheets, i.e. code that defines the site's typography and layout) is not archived, as this could affect the display of the material. These factors are important to keep in mind when viewing archived versions of web materials. Once again, this underlines the point that we cannot with any certainty know to what extent what we see in the archive is similar to what was once online on the live web.

3 Existing web collections

3.1 Introduction

This chapter provides an introduction to some of the existing collections of web material. The focus is on archives that Danish researchers have an opportunity to access in one way or another. The sections describe the main aspects of the archives: their collections, strategies, and ways to access the archives, search functionalities and documentation. The main focus is on the Danish web archive Netarkivet and the US-based Internet Archive, which are probably the archives that Danish researchers will most often use.

3.2 Netarkivet

3.2.1 The collection

Netarkivet is the Danish national web archive that "collects and preserves the Danish part of the Internet" (http://netarkivet.dk). Netarkivet is run by the State and University Library and the Royal Library, who are jointly responsible for collecting and making available the material in the archive. Netarkivet was established as the Danish part of the Internet became subject to the Danish Act on Legal Deposit of Published Material, which lays down the legal framework for the collection and preservation of the Danish cultural heritage.

The first time Internet material came to be encompassed by the Legal Deposit Act was in 1997, when the law was amended to state that the legal deposit obligation applied "irrespective of the medium used to make copies" (Section 1, Act no. 423 of 10/06/1997[15]). This meant that digital publications also became subject to the Act insofar as they were static works, including for example PDF files published on the Internet. However, archiving of the Internet did not begin in the way that we see today; this did not occur until 2005, when dynamic Internet materials also came to be covered by the Legal Deposit Act (Act no. 1439 of 22.12.2004).[16] In order to live up to this obligation it was necessary to find a way to archive the Danish part of the Internet, and Netarkivet was founded to meet this goal.

The Danish part of the Internet should be understood as all material on the Internet that is published in Danish or which targets Danes, including the entire top level domain .dk, and pages in Danish on other domains (such as .com, .eu, .nu, etc.). Only publicly-accessible material is encompassed by the Legal Deposit Act – so password-protected content, for example, is not collected unless it is possible for anyone to get a password. In such cases Netarkivet attempts to create a user

[15] https://www.retsinformation.dk/Forms/R0710.aspx?id=85005

[16] https://www.retsinformation.dk/Forms/R0710.aspx?id=11949. See also the remarks to the bill: http://www.pligtaflevering.dk/loven/bemaerkninger.htm.

login, in order to gain access to, for example, news sites that require login. Materials that are not meant for public access, such as e-mail accounts, bank accounts, intranet networks and the like do not fall within the Danish public web, and no attempt is therefore made to collect them (http://netarkivet.dk/til-webstedejere/faq/#faq_pwd).

3.2.2 Strategies

Netarkivet makes use of three different harvesting strategies to harvest the Danish web: 1) broad crawls (snapshot harvesting), 2) selective crawls (selective harvesting), and 3) event crawls (event harvesting) (Schostag & Fønss-Jørgensen, 2012 p. 110). Special harvestings are also undertaken.

1) **Broad crawls** (snapshot harvesting) are an attempt to harvest all relevant domains. For these harvestings, Netarkivet bases its activities on two types of lists of *URL*s. The first is a list of all Danish domains registered with the Danish national domain name registrar *DK Hostmaster*, i.e. all domains on the top-level domain .dk. The second list is called Danica, i.e. records on Denmark and the Danes – in this context to be understood as websites (or parts of websites) that are in Danish, or which in one way or another relate to Denmark and a Danish audience (e.g. domains hosted on IP addresses in Denmark). This list is maintained by Netarkivet's curators, who are constantly on the look-out for new, relevant domains to be added to the list. On Netarkivet's website, users can suggest domains for the Danica list, which will then be harvested if the curators find that they are relevant.

Broad crawls of the entire Danish part of the Internet was originally envisaged to take about three months, and so the target set in the beginning was to complete broad crawls per year. In the early years this was not possible, however, as several harvestings took longer than expected, due to various circumstances and technical problems.[17] The first broad crawl that Netarkivet undertook thus lasted almost six months (Schostag & Fønss-Jørgensen, 2012 s. 111), so there were not four annual harvestings in the first years. Broad crawls tend to involve more and more material over time, but as the technology improves, it generally takes shorter and shorter time to perform the crawl. Broad crawls attempt to collect as much material as possible, so the harvesting is both wide (on all relevant sites) and deep (encompassing as much of each site as possible, and harvesting up to 25 levels) (http://netarkivet.dk/om-netarkivet/tvaersnitshostninger/).

[17] I will not go into detail here about these challenges, but if you are interested in reading about how various harvestings have been undertaken – which can be useful to know if you wish to use data from a particular harvesting in your research – Netarkivet's newsletters (http://netarkivet.dk/om-netarkivet/nyhedsbreve/) and news archive (http://netarkivet.dk/arkiv/nyhed/) are good sources.

The number of levels indicates how many times the harvester follows links, starting from the URL where the harvesting was commenced. In order to exert the best possible control over the harvesting, the harvesting is in practice undertaken in several steps in which domains up to a certain size are harvested first (this size has changed over the years). Those domains in the first step that have turned out to be larger than the limit are then harvested, up to a new limit, and finally the largest domains are harvested, if possible in their entirety (ibid.).

2) **Selective crawls** (selective harvesting), as the name indicates, are targeted at selected domains. In Netarkivet's selective crawls, around 80-100 highly dynamic domains are selected, i.e. domains with a high level of activity that are considered to be particularly important, such as news sites and "frequently-visited websites belonging to the authorities, the commercial sector and civil society" (http://netarkivet.dk/om-netarkivet/selektive-hostninger/), as well as a smaller number of websites that are considered "particularly distinctive, experimental or unique (e.g. open discussion fora, web communities, personal pages or art sites)" (ibid.). The URLs included in the selective crawls are collected very frequently due to their highly dynamic nature – up to six times a day, and on many levels (Netarkivet, 2014a). On Netarkivet's website, a list can be seen of the URLs (ibid.), and here, too, people can suggest websites for harvesting.

3) **Event crawls** (event harvesting) collect websites containing content about significant events, including planned events, such as elections, major sporting events, the Eurovision Song Contest and the like, and unpredictable events such as natural disasters or man-made crises. There are typically two or three event crawls per year, but it depends on which events are considered relevant to collect. Here, too, harvesting is performed in depth, i.e. on many levels. Netarkivet's website provides an overview of the event crawls (http://netarkivet.dk/om-netarkivet/begivenhedshostninger/).

In addition to these three types of harvests, Netarkivet also undertakes so-called special harvestings, usually for short periods, or on just a single occasion. This might be done if websites are due to be closed, or if a researcher has specific archiving requests in connection with a research project. Some of these special harvestings aim to collect videos (which are not covered by broad or selective crawls). There are for example several special harvestings that include YouTube videos (see http://netarkivet.dk/om-netarkivet/specialhostninger/ for a list).

Special harvestings can also be used to test new technologies for the collection of streamed content, or other content that requires special techniques in order to be collected (ibid.). In addition, there may be other initiatives relating to new ways of collecting material, for example Netarkivet has produced a two-hour video from the virtual world Second Life (http://netarkivet.dk/netarkivet-arkiverer-second-life/).

The amount of data in Netarkivet is growing rapidly, and as of 15.11.2015 it comprised 654 Terabytes (http://netarkivet.dk/om-netarkivet/statistik/). Netarkivet currently uses a technique called deduplication to reduce the amount of data, which should also be mentioned here, because it influences what is archived. Deduplication is designed to automatically delete redundant data, i.e. objects that have already been harvested in previous harvesting operations, in order to save space in the archive. If a picture (such as a jpeg file) has already been archived, and the harvester collects the same picture (with the same URL) again, deduplication will ensure that the image file is not stored in the archive for a second time. Instead, a reference is inserted to the version of the image file that is already stored in the archive databases. Netarkivet calculates that deduplication has brought about a reduction of 50-70% in the number of archived bits in some harvestings (Schostag & Fønss-Jørgensen, 2012 p. 114).

3.2.3 Access

Access to Netarkivet is free of charge, but highly restricted, as, due to the personal information it contains, it may only be used for scientific purposes. At the present time, access is therefore reserved for researchers and PhD students who can obtain online access from home, and for Master's thesis students, who can access the archive from the computers at the State and University Library and the Royal Library. Access must be applied for, and is granted only in relation to a specific research project (or in the case of master thesis students, scientific studies in connection with their Master's thesis project). The applicant form can be found here:
http://netarkivet.dk/wp-content/uploads/2016/02/ansoegererklaering-feb2016.pdf.

3.2.4 Search

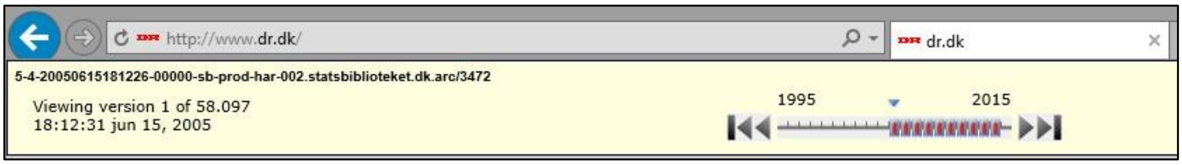
Netarkivet offers two ways of accessing the archive: URL based search and free text search. The URL search takes place in Open Wayback, which is the newest Wayback Machine prototype. As previously mentioned Wayback Machine is a software that can be used to display archived web pages. The user enters a URL into the Open Wayback, after which a results list is shown in the form of a calendar view of all the archived versions of this URL, sorted by the time of harvesting. An asterisk after the date of harvesting in the calendar view indicates that the page has changed, in relation to the previous harvesting (Netarkivet, 2015), but this may reflect minor changes or entirely new content, so it can be difficult to use it as a guideline. Often it may be appropriate just to check out a few pages to see what has changed from one version to another. Dynamic elements, such as a box that constantly switches between three different pictures, also appear to result in an asterisk appearing in the calendar view, even though both versions actually contain the same three pictures.

On the basis of the calendar view, you then select the version you wish to see, and when you click on this the archived version opens in a sort of browser window within the browser. Here, the Wayback Machine performs a kind of replay of a web page (to the extent that this is possible) in order to assemble the archived versions of the various objects of which the page is composed. Pages are thus not archived as whole entities, but rather as many separate objects, which are then reassembled on the basis of the HTML code (and other programming languages) that initially created the page and made it appear to the viewer as a single, coherent object. However, for the sake of clarity, we will speak of an archived version of a web page (with a specific URL), which should be understood to be a page consisting of (and generated by) multiple objects.

As described previously, the Wayback Machine preserves the hyperlink structure of web materials by rewriting links to the archived resources. This means that when you view an archived version of a web page, you can move between archived versions via links, in the same way as when you navigate the live web. However, as mentioned, not everything is archived, and consequently it can happen that there are links to content that is not available in the archive. The Wayback Machine seeks to solve this potential problem by always offering a version of the missing content, if one is available. If, for example, you are viewing a web page which has been archived on a particular day, and you wish to follow a link and navigate to some other content (either on the same page or another page, or on a different site at a different domain), you should optimally, to ensure consistency, be able to view content from the same archiving date. But as this content has not necessarily been archived, the Wayback Machine shows the archived version of the page, which is closest in time to the archiving date of the page containing the link (Internet Archive FAQ, 2014). Alternatively, the error message “Not in archive” is displayed.

The fact that the interface can show the version closest in time may seem a handy feature, but it also creates some new – and certainly not insignificant – complications for researchers who use the archive in the context of research. It means that you cannot assume that linked pages have been archived at the same time, and so the user must, first of all, be very aware of the possibility of encountering temporal ‘leaps’, and, secondly, take into account what this kind of “atemporal linking”(Rogers, 2013 p. 76) may entail for the use of the archived material as a research object. In some cases there may be years or months between the archived versions, so the question is whether one can rightly view these versions as being part of a unit or having a relation or connection. There is certainly a major risk that these pages have never looked like they do in the archive at the same time, and that what the user therefore experiences as different parts of the same website may never have existed simultaneously on the live web.

No special warning is given of such temporal leaps, as it is an integral part of the way the interface works. The time at which a version has been archived is indicated by a yellow top bar at the top of the browser window, where users can always keep themselves informed of the archiving date of the URL that is being viewed. Note that all times are shown as UTC (cf. http://da.wikipedia.org/wiki/UTC), which is one or two hours ahead of Danish time (depending on whether it is winter or summer time) (Netarkivet user manual, Netarkivet (2015)). This is important to know and keep in mind, if you are looking for something that occurs at a very specific time. If, for example, you wish to examine what was happening on a website when a particular television programme was broadcast, it is important that you work out the right harvesting time, which of course is not the same as Danish time.



http://www.dr.dk/

dr.dk

5-4-20050615181226-00000-sb-prod-har-002.statsbiblioteket.dk.arc/3472

Viewing version 1 of 58.097
18:12:31 jun 15, 2005

The top bar in Netarkivet’s Wayback Machine, shown here with an archived version of dr.dk: http://www.dr.dk 5-4-20050615181226-00000-sb-prod-har-002.statsbiblioteket.dk.arc/3472 (18:12:31 June 15, 2005 in UTC time).

The timeline on the left can be used to move back and forth in time – although its functionality in Netarkivet is not optimal, as the timeline is very general, and it is hard to see where you are jumping to (as opposed to the Internet Archive – see 3.3.4 on searching the Internet Archive).

As URLs are crucial to this way of searching the archive, it can be a major challenge for the user if a site has changed its URL. When you search for a URL, you are not informed of whether the website that now occupies this URL previously had a different URL. So unless you know the previous URL, you might erroneously think that the website had not previously been archived, when in fact it might have been archived under a different URL. If you suspect that this could be the case, a potential solution might be to try to find archived versions of concurrent websites in the archive which link to it, and then follow the links from there.

The free text search takes place in the interface Netsearch/Blacklight. You can choose to search for text, URL/domain, links or in all fields. If you search in all fields for a common word, the search will probably return millions or at least hundreds of thousands of results. So it is a very good idea to try to be as precise as possible when you search, which is also recommended in the user manual (Netarkivet, 2015). For a broad search you might have to wait some seconds (the manual mentions up till 10 seconds (ibid.)) before the result is displayed due to the huge amount of material that has to be processed. When you have a result you can use the following facets to limit your search: crawl year, domain, content type (i.e. html, text, image, audio, pdf, etc.) and public suffix (i.e. dk, com, org, info, eu, etc.). You can also see how the results are distributed among the facets, for instance how many of the results are on specific domains. This can help make it easier to use the facets to reduce the number of results. The free text search is a great option but using it can easily result in data overload.

3.2.5 Documentation

As described in the section on the characteristics of archived materials (2.5), the archived web must be regarded as a reconstructed version. This creates a further challenge, as at the time of writing very little documentation is available in web archives, including Netarkivet. The only documentation available to users of the archive that is currently accessible in the results list (calendar view) is a record of how many times a URL has been harvested in the individual years (or rather two-year periods, as the results are sorted in columns of two years), and when this has occurred. In the individual version, the only documentation is the metadata, consisting of the provenance code and the time of harvesting. The provenance code appears as a mouseover box next to the link in the results list (together with the date and time). Few users will be able to use the provenance code other than to see the harvesting time, which is already included in the link and the yellow box of each archived version (see the illustration above). The mouseover function also activates the display of the name of the *WARC* (Web Archive file format) archive file in the bar at the bottom of the browser window. This is shown because it is what the link leads to, just as you can usually can see the URL of a link at the bottom of the browser window. The WARC name consists of: job no.-harvesting time-date-server (Netarkivet user manual, Netarkivet (2015)), but it is not something that can easily be used by the general user.

Another source of a form of documentation is the newsletters and news archive of Netarkivet,[18] and the data on Netarkivet's website about the various harvestings (see the links in 3.2.2). Here you can for example see when various harvesting operations have taken place. Netarkivet's staff has also provided various forms of documentation, including in the form of wikis, which are not currently available to users, but which contain much relevant information on the progress of harvestings and the choices that have been made in this connection.

It is currently being discussed whether more metadata should be made accessible in the future, as this is something that several researchers request because increased knowledge about which harvesting has resulted in a given archived version could help to explain gaps or other problems encountered in the archived material. As there may be many uncertainties associated with the archived materials (see 2.5), it is important that the researcher can obtain as much

[18] http://netarkivet.dk/om-netarkivet/nyhedsbreve/#newsletters and http://netarkivet.dk/arkiv/nyhed/ (in Danish)

documentation as possible, in order to evaluate the material as a research object – and so as to be able to document any claims about gaps or the like in the materials. The development of new tools to make calculations on the basis of various forms of documentation, e.g. in relation to finding the harvestings of a website that have collected the most objects, could also be of assistance to researchers in assessing the status of the archive objects and comparing their relevance.

It could be relevant for researchers to obtain access to some of the automated documentation that is generated during harvesting in the harvester's so-called crawl logs, documenting how a harvesting has proceeded (e.g. whether it has been interrupted along the way). There may also be types of manually-collected documentation that could be useful for researchers to know about – such as the comments of the curators and technicians concerning certain harvesting processes and their progress. Some of the documentation would however be difficult to understand (or perhaps even unintelligible) for the ordinary user of the archive, and a large amount of work would therefore be required to determine how different types of documentation could be made available in an appropriate form.

In relation to documentation, albeit of a different kind, it is also relevant to mention that all searches in Netarkivet are logged (due to the requirements of the Danish Data Protection Agency).

3.3 The Internet Archive

3.3.1 The collection

Another archive which is very useful for researchers in Denmark and around the world is the Internet Archive, which is believed to contain the world's largest collection of archived web materials. The Internet Archive is run by a US non-profit organisation which since 1996 has attempted to archive the entire public part of the Internet, together with other digital sources, with the aim of creating a huge digital library, all of which can be accessed online. Unlike Netarkivet, the Internet Archive is not based on national legislation regarding legal deposits.

At the time of writing, the Internet Archive encompasses over 491 billion pages (https://archive.org), and the archive collects about one billion pages per week (Kahle, 2015). It contains websites in more than 40 languages, including a large amount of Danish material (http://netpreserve.org/archives/internet-archive).

The Internet Archive also archives many other types of digital content besides websites. At the time of writing, the archive contains more than 10 million texts, almost 3 million audio recordings, more than 2.5 million videos, more than 160,000 recordings of live concerts and other content such as images and software (https://archive.org).

3.3.2 Strategies

The Internet Archive does not disclose specific information on their website about their archiving strategies, and there does not appear to have been a consistent strategy over the years. The first harvestings were a form of thematic harvesting as the websites harvested were those relating to the presidential elections in 1996, and later, in collaboration with the Library of Congress, the elections of 2000 and 2002 (Kimpton & Dubois, 2006). The Internet Archive soon partnered with Alexa Internet, a commercial company that began harvesting the web in 1996 in order to obtain data for one of its products: a toolbar for browsers (ibid.). For Alexa Internet, what was primarily relevant was the user data that the harvester downloaded (e.g. which sites were most frequently visited and which sites they could recommend as related), and not so much the actual archived websites, so they passed on the harvested data to the Internet Archive, which could use them to display archived versions of websites (ibid.).[19]

Today, the archive harvests data in many different ways, both as very broad crawls and as selective and thematic crawls, etc. The archive harvests data on the basis of accumulated lists, in which the links they encounter along the way are continuously added to the lists that direct the archiving. The archive also receives donations of data from various places, so it is a highly heterogeneous and composite collection. Compared to Netarkivet, the quality of the archived material in the Internet Archive is often less good. The Internet Archive frequently does not archive in as much depth as Netarkivet's harvesters, and in some cases only the top layer of a website is preserved.

3.3.3 Access

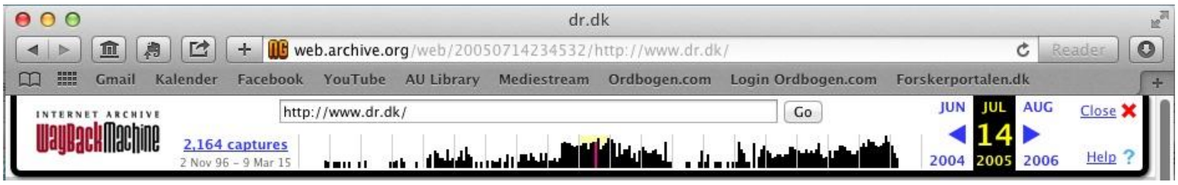
Access to the Internet Archive is free of charge for all at: https://archive.org, where one can access both the Internet Archive's Wayback Machine and the other collections of digital materials.

3.3.4 Search possibilities

At present, the material can be accessed both via thematic/topic-based collections and through URL searches in the archive, which is what we will be focusing on here. The Internet Archive is accessed via the user interface Wayback Machine. The idea is to give the user the same experience of navigating the web as you would obtain from the live web – but instead at a selected historical time. Originally, part of the point was also to get rid of all the broken links, which were a great inconvenience when navigating around the web (see Kimpton & Ubois (2006) for a description of the Internet Archive's first year and some of the ideas about the archive).

[19] See also Masanes (2005) for information on the Internet Archive's co-operation with Alexa Internet.

At the moment it is only possible to search for URLs in the web archive, and the same interface is used as in Netarkivet, namely the Wayback Machine. It has a slightly different appearance, however, as Netarkivet has made some modifications to the version it uses. The results list looks different, but the main difference (at least from a user's perspective) may be seen in the way the timeline is composed.


web.archive.org/web/20050714234532/http://www.dr.dk/

http://www.dr.dk/

Front page, dr.dk_ 14 July 2005 [http://www.dr.dk] _Internet Archive_ [http://web.archive.org/web/20050714234532/http://www.dr.dk/].

While the timeline in Netarkivet is at a very general level, in the timeline of the Internet Archive it is easier to use to move between archived versions by skipping forward or backwards in time. If you use the dates indicated, you can for example jump to the nearest archived version before or after by clicking on the blue arrows (which are greyed out if there is no earlier or later version), or to the nearest version in the month before or after by clicking on the specified month, or to the year before or after by clicking on the year.

Another way to jump between archived versions in the Internet Archive via the timeline is to use the white boxes with black bars. Each black bar represents a month, and the length of the bar illustrates the number of times the URL has been harvested in the month in question. The longer the bar, the more captures from the month in question. The bar is clickable, allowing the user to move to the nearest archived version in that month.

Note also that the URL that is archived appears in the lower address field, which can also be used to enter a new URL to search for (using 'Go'), while the URL that the archived version has been given is shown in the browser's normal address field (see the illustration above). In contrast to Netarkivet, the Internet Archive is accessible online on the live web, and each archived version of a URL has thus been given a new URL, which is the live web address where the archived version can be found (see also the documentation below).

3.3.5 Documentation

The documentation available in the Internet Archive is limited to the Wayback Machine's record of how many times a URL has been crawled, and when. In the individual archived versions you can see what the URL is, how many times it has been crawled, during which period, and on what date (but not at what time) the version has been harvested. In the browser window's address bar, you can also see the URL that has been assigned to the resource (i.e. the archived version). When the archive is online, this URL can always be used to find the version again. This URL is also important as documentation, because it contains the exact harvesting time, which is not otherwise visible on the page.

Structure of URLs in the Internet Archive's Wayback Machine:

https://web.archive.org/web/20150104211940/http://www.dr.dk/
Wayback Machine URL/collection/ time shown as yyyymmddhhmmss/URL

Another source that can be used as documentation more generally, in relation to how the archive and the Wayback Machine work, is the Internet Archive's FAQ. The part dealing with the Wayback Machine is particularly relevant for researchers (https://archive.org/about/faqs.php#18).

3.4 Other national web archives

A brief mention should be made of a number of other web archives that it may be useful for Danish researchers to know about. As they are all available online, they can be accessed by Danish researchers, and you can consequently be lucky enough to find material here that can supplement material in the Danish archives. This may also be highly relevant when you are looking for non-Danish websites. The following descriptions are mainly based on information from the International Internet Preservation Consortium, which has a very useful list of 34 web archives that are members of the consortium. The list may be viewed here: http://netpreserve.org/resources/member-archives. There is also a timeline of when the various archives were created: http://viewshare.org/views/abpo/iipc-member-archives-2/. Another useful source is the so-called environmental scan made by Truman (2016) as a Harvard Library report, which "document web archiving programs from 23 institutions from around the world and report on researcher use of – and impediments to working with – web archives." (Truman, 2016, p. 3).

3.4.1 Library of Congress Web Archives

The US Library of Congress Web Archives (LCWA) collect and make available collections of digital material, including several collections of websites selected by specialists to cover specific themes and topics that may be relevant to researchers. The Library of Congress Web Archives undertakes selective crawls, event crawls, and thematic crawls, and since 2000 they have collected and preserved collections of relevant websites in connection with events such as elections in the United States, the war in Iraq and 11 September 2001 (http://loc.gov/webarchiving/). The archive is well documented and curated with documentation about both collections and each website. The archives may be accessed online, and there are multiple search options: search by URL, faceted search, browsing (alphabetically or by subject) or search in current collections (http://netpreserve.org/member-organizations/library-congress).

3.4.2 The UK Web Archive

The UK Web Archive has two main collections. One of these is a collection of more than 5,000 sites of cultural, political, social and historical significance, selected by leading institutions in the UK (http://netpreserve.org/resources/member-archives).

> “The UK Web Archive contains websites that publish research, that reflect the diversity of lives, interests and activities throughout the UK, and demonstrate web innovation.” (http://www.webarchive.org.uk/ukwa/info/about)

The material in this collection is harvested in selective, thematic and event crawls. The archive, which has existed since 2005, offers several interesting ways of searching the materials. You can search by website title or URL, but there is also the possibility of full text searching and browsing (alphabetically, by subject or in special collections). This collection of websites may be accessed free of charge at http://www.webarchive.org.uk/ukwa/.

In 2013, the British Library, which operates the UK Web Archive, began to archive the entire UK web domain under British legal deposit legislation (Non-Print Legal Deposit Regulations 2013). The archiving in this collection is done in co-operation with the other legal deposit libraries in Britain and Ireland: the Scottish National Library, the Welsh National Library, Cambridge University Library, the Bodleian Library in Oxford and the library of Trinity College Dublin. However, these do not offer online access, as access to this archive is only granted locally at the six legal deposit libraries
(http://www.webarchive.org.uk/ukwa/info/about).

3.4.3 Pandora

Another archive that was established at almost the same time as the Internet Archive in 1996 is the Australian Web Archive PANDORA. Unlike the Internet Archive, PANDORA does not attempt to archive the entire Internet (and other digital materials); the strategy here is to undertake selective crawls and event crawls, so that the archive covers a variety of topics relating to Australia and Australians. The archive is freely accessible online, and may be searched in several ways: by URL or keyword, browsing (alphabetically, by subject) and by free text search. The archive may be accessed at http://pandora.nla.gov.au.

3.4.4 The Portuguese Web Archive

The Portuguese Web Archive is the Portuguese national web archive, and it is mentioned here because it is available online in its full extent and with the possibility of full text search. This contrasts with other national web archives in the public sector which do not provide full online access to the archives for all. The archive has been harvesting the Portuguese web since 1996, and since 2007 has been run by the Foundation for National Scientific Computing (FCCN) in Portugal. The archive may be accessed free of charge at http://www.archive.pt.

4 The researcher's own web archive

4.1 Introduction

When, as a researcher, you attempt to compile your own archive of web material, it is important to start by deciding precisely what it is you wish to archive, and why. What is the material to be used for, and by extension, what methods will be most useful for this purpose? If you do not determine precisely what research question the material is intended to help you answer, and how you intend to analyse the material, there is a risk that it may not be possible to use the collected material for the intended purpose (Brügger, 2005 appendix 2 s. ii). It can also be useful to consider what other methods can be combined with the archiving and analysis of the archived material in order to strengthen the study, so that, as far as possible, you form a clear idea of what the archived web material can contribute to, as well as what it will not be able to say anything useful about.

Once you have decided what it is you wish to analyse and how, the next step is to think in concrete terms about how the various types of web archiving can contribute to the construction of a research object that can be used in the analysis. In order to decide on one or more methods, I recommend that you read the previously mentioned publication "Methods of collecting facebook material and their effects on later analyses" (Laursen, Brügger & Sandvik, 2013), which offers valuable insights in the advantages and disadvantages of different methods. I also recommend Brügger's *Archiving Websites : General Considerations and Strategies* (Brügger, 2005), which includes a detailed description of strategies for creating your own web archive. This publication offers insights into important aspects as, among other things, the dynamics of web, the archiving strategies and software that can be applied, and the differences between these strategies (and software) Note also that appendix two contains a step-by-step guide to archiving a website, as well as some tables that can serve as aids when choosing archiving strategies and tools. These can usefully be applied in conjunction with the points from this publication. Of course I also hope that the publication that you are currently reading will be helpful as I try to gather relevant points from the abovementioned publications as well as other sources.

When you archive web materials yourself, it is important to ensure that you document what you have done, because, as described above, this has an influence on the kind of research object you will end up with. Before you begin archiving, it is a good idea to plan exactly what you want to archive, when and how. Brügger recommends that you always draw a site map in order to visualise and illustrate the structure of the website (or parts of it) that you wish to archive (Brügger, 2005 appendix 2 p. vii). You can also use the site map in relation to an archiving log, where you note which parts you have archived, how you have done it, and any problems that may have arisen in the process. Depending on how much you archive, it can often be difficult to remember the individual parts of the process afterwards, and a log can therefore be a very good tool to remind you of what succeeded or failed, and why, etc.

I would also recommend that you give some thought at the outset to how you intend to structure your archives – how you will name the files, for example, and what system you will use to keep track of them. Documentation can either be in a separate document in the form of a log, or you can use a data management program in which you have the option of adding metadata.

Finally, it is important to be aware that many of the problems that are described in relation to macro archiving in the section on the characteristics of archived material (2.5) will also apply to the archived versions that you end up when you create your own archive.

The following reviews some of the methods that can be used for micro archiving. However, it should be noted that some of these methods, including screen recording, are also sometimes used by libraries to archive certain materials (cf. Netarkivet's special harvestings).

4.2 Various methods for creating your own archive

4.2.1 Screen capture

One of the easiest and most immediately accessible ways to archive web material is to take a screenshot of a web page. Most computers have a program that can take a picture of what appears on the screen – either of a selected area, a selected window, or the entire screen. Taking a screenshot gives you a version that is similar in appearance to the one that was on the screen, but only the selected part, and only that which can be captured in a picture – any dynamic elements will not be included. The type of file you use to save the screenshot is very important. Some programs simply take a snapshot and thereby convert the HTML code from which the page is built up into a graphic file (such as jpeg, png or tiff).[20] This means that any links in the page will no longer work.

However, there are also applications such as Web Snapper that take what looks like a screenshot, but the whole page – if it is longer than what can be seen on the screen at one time – is included, and the HTML is preserved, so that the links will still be functional. Once the page has been downloaded, the image of it can be saved as a PDF file, in which the links are still clickable. There are also differences in whether the file that you obtain is searchable or not. Searchability requires that the text in the image can be read by the computer – and this cannot be done if the page is saved in a graphic format (such as jpeg, png or tiff). However, if the page

[20] The page may of course also include code in other languages, such PHP or JavaScript.

is stored as a PDF file in which the text is machine-readable it is searchable. If the text on the website is machine-readable, the PDF file in which Web Snapper saves the archived page will also be machine-readable. If not, for example if the text is included as part of an image (and thus in a file format that is not immediately machine-readable), you can use a program like Acrobat Pro to scan the PDF file so that the image of the text is converted to electronic text using Optical Character Recognition. The text then becomes machine-readable and thereby searchable.

It may be advantageous to try using several methods, such as both a normal screenshot function and Web Snapper, because there may be things that can be obtained with one version and not with another, so the best solution can sometimes be to combine methods. This also applies in relation to the methods that will be described in the following; it may often be beneficial to combine for instance screen filming or archiving via API with screenshots because a screenshot preserves the image of the page as you see it at the given moment.

Pros of screen capture:

- A screen capture looks exactly like what you see when you look at the website (without interacting with it), i.e. with all the textual and static elements and layout preserved. (The term ‘text’ should be understood broadly to include images, graphics, etc.)
- You can preserve a chosen portion of the screen, large or small
- Some programs can take a snapshot of an entire web page, even if it is bigger than what you can see at one time in a browser window
- Some programs can preserve links
- Some programs provide an output that is machine-readable and thereby searchable

Cons of screen capture:

- Sound and moving images cannot be archived
- Dynamic content or content that requires user interaction cannot be archived
- Not all programs can save an entire page
- Not all programs can save links
- Not all programs provide screen captures that are machine-readable and searchable

(Laursen, Brügger & Sandvik, 2103)

4.2.2 Screen recording

A screenshot, as described, cannot capture the dynamic elements of a web page, so streamed content or objects that are constantly changing or being updated cannot be captured with the tools used to take screenshots. In order to archive this type of material, it may be useful to use a tool designed for screen recording, such as Snaps Pro X or Snagit, which are capable of capturing materials that change over time. Screen recording can be used to collect both audio and video, and, as with screen captures, you can select a sample of what is displayed on the screen, or the full screen. You can choose to record a screen section without doing anything, or to interact with the content on the screen while you record (if you are present during the screen recording). If you so choose, you can scroll up and down a page that is larger than the camera can capture in a single shot, and thereby record more of the page. You can also interact with objects or follow links, and thereby move beyond the page itself. Some programs, such as Snaps Pro X, allow for different settings, such as recording at different frame rates, or choosing between a fixed camera, panning or a camera that follows the cursor.

In relation to the previously mentioned project by Laursen, Sandvik and Brügger "Cross media production and communication", Ditte Laursen, Bjarne Andersen and Mads Ravn from The State and University Library has developed a program that can be used to schedule and automate screen recording (http://netarkivet.dk/nyt-vaerktoej-til-netarkivet/). With the tool, which is now operational in Netarkivet, particular times can be chosen at which the recording will start and end at a specific URL. The tool can also be programmed to click in certain places and enter something, if it is necessary to log in or the like. At the State and University Library, the tool is used to complement the other harvestings on special occasions, such as events with live streaming online. The program is open source, and so can also be used by other researchers. You can read more about the recorder by following the link mentioned above. See also Laursen, Brügger & Sandvik (2013) for a description of how to use automated reload of a page in order to capture developments over time.

Pros of screen recording:

- What is archived resembles what was online
- Sound and moving images (also in the form of streamed content) can be archived
- Dynamic information and user interaction can be archived
- Development over time can be monitored
- Can be used to show the interrelations between web elements, web pages or websites by showing what happens when links are followed

Cons of screen recording:

- Only a selected part of the website is recorded
- Only the interaction that took place during the recording is preserved
- Not machine readable or searchable
- Link structures are not retained (except insofar as they can be filmed)

(Laursen, Brügger & Sandvik, 2103)

4.2.3 Link crawling (HTTrack)

HTTrack is a program that can archive web material by searching links (link crawling). The program can be used to archive a version of a website by downloading the files to a local folder on your own computer, while preserving the link structures. The version can then be opened in a browser which replays the site from the archived files, including links – which means that you can also navigate around the site. HTTrack thus provides several of the same advantages as the harvestings undertaken by the web archives. However, as it is on a smaller scale, you cannot navigate in the same way as in the web archives, since you can only follow the links that have been archived, and only if the content of the destination (the URL that is linked to) is also archived. On the other hand it is a very good way to preserve an entire website in a coherent form, if you wish to ensure that the material is preserved at a particular time – which you cannot be certain of if you rely on the harvestings of the web archives. You can find HTTrack at: http://www.httrack.com. Other programmes that archive by link crawling are, for instance, SiteSucker for OS X (http://ricks-apps.com) and WARCreate (http://warcreate.com). At present we recommend using HTTrack because we have good experience using this software and it is relatively easy to use once it is installed (cf. NetLab's website for a manual describing how to install HTTrack).

The pros and cons of link crawling are mentioned in 2.2.1 on web archiving via web crawlers. In relation to the researcher's own archiving it is also relevant to mention that the amount of data might be a challenge if many websites are crawled in full. This points to the need for reflections on how to work with the data after the archiving – but then again, it is always better to have the data archived than suddenly realising that they are no longer online!

4.3 On-demand archiving

4.3.1 Netarkivet

Researchers may request Netarkivet to perform a customised harvesting of websites that they need for a research project. Not all types of tasks can be performed within Netarkivet's resources, though, and in the future there is a possibility that you may be required to pay for harvesting that lies outside what Netarkivet already does.

You can always suggest a website to Netarkivet if they are not already harvesting it. Netarkivet, like several other web archives, is pleased to be informed if you know of relevant Danish websites that lie on other top-level domains besides .dk. At http://netarkivet.dk/du-kan-hjaelpe-netarkivet/ you can test whether they know of a domain, and here you can also suggest URLs to be harvested in selective or event harvestings. Netarkivet's curators will then evaluate whether this URL should be included.

4.3.2 Archive-It

A possibility for on-demand archiving is to order an archiving from Archive-It. Archive-It is a commercial company that originated from the Internet Archive, and it is also the Internet Archive that stores the archived material. Archive-It offers a subscription service in which customers can use Archive-It's online platform and software to collect, catalogue, administer and make available collections of digital material, including websites (https://www.archive-it.org/learn-more). Free text search is possible. They also offer Archive-It Research Services (ARS), which provides access to data sets extracted from collections (metadata, link graphs, named entities, other data). At the website https://www.archive-it.org, it is possible to explore some of the collections, including a number of collections from their partners, for instance some event collections.

5 Referencing archived web material (best practice)

In addition to the characteristics of the archived web and the technical challenges relating to various kinds of web archiving that have so far been mentioned in this book, other significant challenges exist for the researcher who wishes to use archived web materials in research. These challenges relate to rights management (data protection, copyright, etc.) which apply to archived material across harvesting methods. Very few people are accustomed to citing and applying archived web material in their research, so we will therefore briefly touch upon a few points in relation to how to use the materials and how to cite them. It is first and foremost important to be aware of the fact that the materials in the web archives are protected by copyright law as they were on the live web, and this must of course be taken into account when using protected material (cf. the citation rules in section 22 of the Copyright Act).[21]

5.1 Material from Netarkivet

However, there are further significant restrictions on how content from Netarkivet may be used, because the content may be protected not just by the Copyright Act, but also by the Act on Processing of Personal Data. At Netarkivet's website, you can read about typical uses of the archived material, and what is permissible in relation to copying, use for teaching, presentations and publications: http://netarkivet.dk/adgang/ (in Danish). A contact e-mail address is also available if you are in doubt and require assistance.

In relation to usage, it is for example permitted to make copies of websites for your own use, which is a great help when you are working with the materials, as it is not currently possible to delimit a corpus in the archive. Since there is so much material in the archive – especially if you are studying one of the large websites that are crawled very often – it can be convenient to create your own micro archive of selected parts of the archived material to make it easier to analyse it.

You are also allowed to show websites in or from the archives for teaching purposes, at scientific conferences, etc., but there are certain requirements that must be met, which are listed on the site (cf. link above). It is always a good idea to check here to keep yourself informed about the current rules.

In relation to publications, the site states that you may:

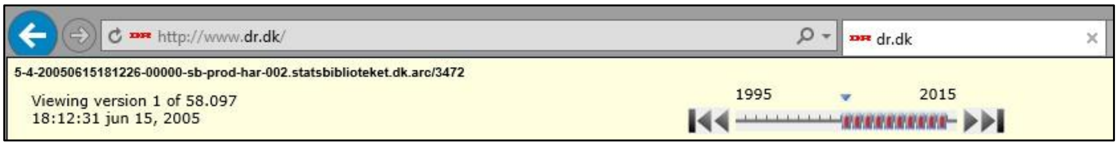
[21] Section 22 of the Copyright Act: "It is permitted to quote from a published work in accordance with good practice and to the extent justified by the purpose." (Act no. 1144 of 10.23.2014). With regard to what is 'good practice', we refer the reader to Hielmcrone (2013), which is available at: https://www.statsbiblioteket.dk/om-statsbiblioteket/filer/de-vigtigste-ophavsretlige-bestemmelser-5.-rev.-udg. See also http://kubis.kb.dk/ophavsret and www.forskerportalen.dk.

> "... reproduce the archived websites in the form of screenshots in critical or scientific presentations, but only in connection with your text, and only to the extent justified by the purpose." (Ibid.)

Please note also that the Act on Processing of Personal Data stipulates that all "personally identifiable information" (ibid.) must always be rendered unidentifiable at showings and in reproductions.

When you make a reference to an archived version of a website in Netarkivet, Netarkivet's Wayback Machine manual recommends that you state the specified URL, the provenance code (see 3.2.5) and the exact harvesting time of the archived version (Netarkivet, 2015). Both the provenance code and the exact harvesting time can be seen in the yellow top bar of the archived version.

Here, for example, is the upper part of the display of the first archived version of www.dr.dk:



http://www.dr.dk/

dr.dk

5-4-20050615181226-00000-sb-prod-har-002.statsbiblioteket.dk.arc/3472

Viewing version 1 of 58.097
18:12:31 jun 15, 2005

According to Netarkivet's recommendation, it should therefore be referenced like this:
http://www.dr.dk 5-4-20050615181226-00000-sb-prod-har-002.statsbiblioteket.dk.arc/3472 (18:12:31 Jun 15, 2005, UTC time).

Note that I have added "UTC time". This is because harvesting times, as previously mentioned, are specified in UTC time in, which means that they are one or two hours ahead of Danish time (depending on whether winter or summer time is in force) (Netarkivet, 2015).

In the results list, the names of the individual archived versions are stated together with their time of harvesting (apart from seconds), and the provenance code for each version can be seen by hovering your mouse over the link. In this way, you or others can find the material on the basis of a reference formulated in the above manner.

5.2 Material from the Internet Archive

The FAQ of the Internet Archive indicates how to reference versions of websites that have been found via the Internet Archive's Wayback Machine. In this connection, the Internet Archive has made inquiries to the MLA (Modern Language Association), which is one of the most common formats in the humanities, according to Purdue OWL.[22] However, the MLA has not yet established an MLA format for how to correctly refer to a source/resource like the Wayback Machine. To be on the safe side, therefore, they suggest that the reference should be as comprehensive as possible:

> "…it's best to err on the side of more information. You should cite the webpage as you would normally, and then give the Wayback Machine information." (Internet Archive FAQ, https://archive.org/about/faqs.php#265)

They also provide the following example of a reference:

> "McDonald, R. C. "Basic Canary Care." _Robirda Online_. 12 Sept. 2004. 18 Dec. 2006 [http://www.robirda.com/cancare.HTML]. _Internet Archive_. [http://web.archive.org/web/20041009202820/http://www.robirda.com/cancare.HTML]." (Internet Archive FAQ, https://archive.org/about/faqs.php#265)

Note that the reference includes both the original URL (the URL that is archived), and the URL of the archived version (which consists of the Internet Archive's URL followed by the year, date and time of archiving of the URL). If no date is given for when it was updated, the MLA suggests that you use "the closest date in the Wayback Machine" (ibid.). In addition, you should also include the date on which the website was 'retrieved' (ibid.). I assume that this means the date on which the material was downloaded from the Internet Archive (just as you would refer to a text that is online). In this format, you do not therefore necessarily need to state the harvesting/archiving date, as this is automatically clear from the URL of the archived version (see above). This URL can also always be used to identify the archived version (see 3.3.5).

5.3 References to your own archive

Similarly, when creating your own archive of web materials, it is important to consider how best to reference the archived material. You could for example base your system on some of the same above-mentioned principles – it is at any rate important to state which URL has been archived, and when. If you do not use a program that organises the files itself (as HTTrack does), but, for example, takes screenshots or screen recordings, it is a good idea to decide from the outset which reference format you wish to use. In this way, you can be sure that you always know the provenance. One possibility could be to decide that the file names of the archived files will always start with the year, month and date (and perhaps also the time, if you consider that to be important) and then specify the URL (possibly with an underscore in between). Such a reference could look like this: 150211_http://www.netlab.dk. Other formats can be used as well – the important thing is to be consistent, so that you can find your way around the archive. But it is certainly an advantage to include the date in the format, as this will make the files easier to sort by date in a systematic manner. Note that the order mentioned

[22] https://owl.english.purdue.edu/owl/resource/747/01/

above (year, month, day) will result in the files being sorted chronologically, contrary to for instance day, month, year. You can also note references in your log if you want to be sure that you will be able to find out which materials have been archived and when, and perhaps also to be able to link this with the methods used.

5.4 Reference rot – another reason to reference the archived web

Besides the fact that it is essential to reference the archived web if you are using it as a source, there can also be advantages in referring to archived rather than live web, even if the material is still is available on the internet – because it is far from certain that it will continue to be there! If archived web material is publicly available, as it is in the Internet Archive, it may be useful to refer to this rather than to the live web, since in this way you can be (almost) sure that the source you refer to will not disappear – which otherwise appears to be a significant risk, according to projects that deal with so-called 'reference rot' (Klein et al., 2013) in academic publications. 'Reference rot' can be used to describe both broken links (links that no longer work, i.e. no longer lead to the appropriate web object) and so-called 'content decay' (see for example http://mementoweb.org/missing-link/), which refers to the fact that the content linked to may change over time due to updates, etc. (see also www.hiberlink.org). Content decay may result in what is actually referenced no longer being shown in the link used, which means that the reference is incorrect – without this necessarily being evident to the person who follows the link. One of the studies that deals with reference rot is that of Klein et al. (2013), who studied more than one million references to web resources in more than 3.5 million articles in STM (Science, Technology and Medicine). The study showed that seven out of ten articles containing references to web resources suffered from reference rot.

6 Glossary

- Bot trap: A ‘trap’ (intended or unintended) for the harvester that generate links and creates an endless loop of requests, causing the harvester to go in circles or crash.

- The live web: The web that is online, in contrast to the archived web. The reason for using this term rather than ‘the online web’ is that the latter term creates confusion in relation to those archives, such as the Internet Archive, that actually have their files available online. Material from these archives is thus at one and the same time an archived version of the live web, and actually present on the live web. However, in most cases, it is useful to distinguish between the archived and the live web.

- DK Hostmaster: The Danish domain name registrar and administrator of all domains in the top level .dk domain.

- Domain (web domain): A name that can be registered in a TLD (top-level domain), such dr.dk, tv2.dk, etc. The domain has an IP number. The domain may correspond to a website, but not necessarily. A website may span multiple domains, while a single domain may host several websites.

- HTTP (Hyper Text Transfer Protocol) is the protocol used on the World Wide Web. HTTPS, which stands for Hyper Text Transfer Protocol Secure, is used for secure connections.

- URL: Uniform Resource Locator. URL is a standard for how to describe an address on the web. A URL is, thus, a web address in a particular format, i.e. a way of describing the address of a particular resource on the Internet (in the same way as when you write the address of a person, you write their name, street name, house number, postcode and possibly country).

The structure of URLs on the World Wide Web (www):

protocol://subdomain.domain.top-level domain/path/page/

http://dac.au.dk/forskning/forskningsprogrammer/

- Not all URLs have subdomains and paths/pages. Another standard for how to describe a web address is URI (Uniform Resource Identifier), which some people prefer to use. The web address is the same, irrespective of whether you call it a URL or a URI.

- Link crawling (also called web crawling): harvesting by means of a so-called crawler that explores links and archives files. See 2.2.1.

- Robots.txt exclusion: A de facto standard that can be added at the root of a domain, which instructs automated systems not to crawl a site or parts of it. See 2.2.1.

- Top-level domain: The domain that is at the end of the URL, e.g. .dk, .com or .org. These may be country domains or generic domains. There are also subdomains, such as in the UK, which has co.uk, ac.uk, etc.

- WARC file format: A compressed file (like a zip file) containing web resources such as pictures, HTML pages, style sheets, etc. + some metadata. In the archives, they have a size of 100 MB.

- Web bot: A software application that performs automated tasks on the web, and which can be used for many types of tasks, including indexing, updating and archiving.

- Web element: The smallest meaningful unit of a web page (according to Brügger, 2009). See 1.3.

- Web page: What can be seen in a browser window (and has a URL). See 1.3.

- Website: Several web pages that are linked together in formal, semantic and physically performative terms (cf. Brügger, 2009). See 1.3.

Other glossaries of web archiving terms:
http://www2.archivists.org/glossary#.V3Yo_VfArdk
https://webarchive.jira.com/wiki/display/ARIH/Glossary+of+Web+Archiving+Terms

7 Bibliography

Brügger, N. (2001). The last page of the internet? The Importance of Preserving the Dynamic Aspects of the Internet. Presented at the Preserving the Present for the Future - Strategies for the Internet. Retrieved from http://www.deflink.dk/arkiv/dokumenter2.asp?id=695

Brügger, N. (2005). *Archiving Websites*. Aarhus: Centre for Internet Research:

Brügger, N. (2008). The Archived Website and Website Philology : A New Type of Historical Document? *Nordicom Review*, *29* (2), 155–175.

Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society*, *11* (1-2), 115–132.

Brügger, N. (2010). Introduction: Web History, an Emerging Field of Study. In N. Brügger, *Web history* (pp. 1–25). New York: Peter Lang.

Brügger, N. (2011). Web Archiving - Between Past, Present, and Future. In M. Consalvo & C. Ess, *The Handbook of Internet Studies*. John Wiley & Sons.

Brügger, N. (2012). When the Present Web is Later the Past: Web Historiography, Digital History, and Internet Studies. *Historical Social Research/Historische Sozialforschung*, *37* (4), 102–117.

Brügger, N. (2013a). Facebooks historie. Udviklingen af en tom struktur. In J. L. Jensen & J. Tække, *Facebook : fra socialt netværk til metamedie* (pp. 17–42). Frederiksberg: Forlaget Samfundslitteratur.

Brügger, N. (2013b). Historical Network Analysis of the Web. *Social Science Computer Review*, *31* (3), 306–321. doi:10.1177/0894439312454267

Brügger, N., & Finnemann, N. O. (2013). The Web and Digital Humanities: Theoretical and Methodological Concerns. *Journal of Broadcasting & Electronic Media*, *57*(1), 66–80. doi:10.1080/08838151.2012.761699

Day, M. (2006). The Long-Term Preservation of Web Content. In J. Masanes, *Web Archiving* (pp. 177–199). Springer.

Dougherty, M., & Schneider, S. M. (2011). Web Historiography and the Emergence of New Archival Forms. In D. W. Park, N. W. Jankowski, & S. Jones, *The long history of new media : technology, historiography, and contextualizing newness* (pp. 253–266). New York: Peter Lang.

Dougherty, M., Meyer, E. T., Madsen, C. M., Van den Heuvel, C., Thomas, A., & Wyatt, S. (2010). *Researcher Engagement with Web Archives: State of the Art*.

Truman, G. (2016). Web Archiving Environmental Scan. Harvard Library Report.

Hielmcrone, H. (2013). De vigtigste ophavsretlige bestemmelser : Kort vejledning for biblioteker. 5. reviderede udgave. Available at https://www.statsbiblioteket.dk/om-statsbiblioteket/filer/de-vigtigste-ophavsretlige-bestemmelser-5.-rev.-udg

Internet Archive FAQ, https://archive.org/about/faqs.php - The_Wayback_Machine.

Kahle, B. (2015, February 11). Locking the Web Open, a Call for a Distributed Web | Internet Archive Blogs. *Blog.Archive.org*. Retrieved March 5, 2015, from http://blog.archive.org/2015/02/11/locking-the-web-open-a-call-for-a-distributed-web/

Kimpton, M., & Dubois, J. (2006). Year-by-Year: From an Archive of the Internet to an Archive on the Internet. In J. Masanes, *Web Archiving* (pp. 201–212). Springer.

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2013). Scholarly context not found: one in five articles suffers from reference rot. *PLoS ONE*, *9*(12), e115253–e115253. doi:10.1371/journal.pone.0115253

Laursen, D, Brügger, N. & Sandvik, K. (2013). Methods for collecting Facebook data and their effects on later analysis. Paper presented at NORDMEDIA.

Library of Congress. (n.d.). Michele Kimpton - Digital Preservation (Library of Congress). *Digitalpreservation.Gov*. Retrieved March 30, 2015, from http://www.digitalpreservation.gov/series/pioneers/kimpton.HTML

Lyman, P. (2002). Archiving the World Wide Web. *Clir.org*. Council on Library and Information Resources and Library of Congress. Retrieved July 29, 2014, from http://www.clir.org/pubs/reports/pub106/web.HTML

Masanes, J. (2002). Towards Continuous Web Archiving. *D-Lib Magazine*, *8* (12). doi:10.1045/december2002-masanes

Masanes, J. (2005). Web Archiving Methods and Approaches: A Comparative Study. *Library Trends*, *54* (1), 72–90. doi:10.1353/lib.2006.0005

Masanes, J. (2006). Web Archiving: Issues and Methods. In J. Masanes, *Web Archiving* (pp. 1–53). Springer.

Netarkivet (2015). Brugermanual til Netarkivet. The State and University Library and The Royal Library.

Nielsen, J. (2014). DR's undervisning på tværs af medier: En historisk undersøgelse af mediesamspil. PhD thesis submitted at the Department of Aesthetics and Communication, Aarhus University.

Pligtaflevering.dk. (n.d.). Bemærkninger til lovforslaget. *Pligtaflevering.Dk*. Retrieved February 22, 2013, from http://www.pligtaflevering.dk/loven/bemaerkninger.htm

Rogers, R. (2013). The Website as Archived Object. In R. Rogers, *Digital Methods* (pp. 61–82). Cambridge, MA: MIT Press.

Schneider, S. M., & Foot, K. A. (2004). The Web as an Object of Study. *New Media Society*, *6* (1), 114–122. doi:10.1177/1461444804039912

Schneider, S. M., Foot, K. A., & Wouters, P. (2010). Web Archiving as e-Research. In N. W. Jankowski, *e-Research : Transformation in Scholarly Practice* (pp. 205–221). New York, London: Routledge.

Schostag, S., & Fønss-Jørgensen, E. (2012). Web archiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective. *Microform & Digitization Review*, *41*(3-4), 110–120. doi:10.1515/mir-2012-0018

Taylor, N. (2012, October 26). Using Wayback Machine for Research | The Signal: Digital Preservation. (B. Lazorchak)*Blogs.Loc.Gov*. Retrieved February 27, 2015, from http://blogs.loc.gov/digitalpreservation/2012/10/10950/

Thomas, A., Meyer, E. T., Dougherty, M., Van den Heuvel, C., Madsen, C. M., & Wyatt, S. (2010). *Researcher Engagement with Web Archives: Challenges and Opportunities for Investment*. *JISC*. Joint Information Systems Committee.