2

Web Archiving – Between Past, Present, and Future

Niels Brügger

This chapter argues that one of the most important issues for Internet scholars in the future is to get a hold on the Internet of the past – and the Internet of the present. There are two reasons for this. First, the archiving of the web enables us to write web *history*, which is a necessary condition for the understanding of the Internet of the present as well as of new, emerging Internet forms. Second, it enables us to document our findings when we study *today's* web, since in practice most web studies preserve the web in order to have a stable object to study and refer to when the analysis is to be documented (except for studies of the live web). Therefore, the problems related to web archiving as well as the special characteristics of the archived web document are not only important to web historians; they are also of relevance for any web scholar studying, for example, online communities, online games, news websites, etc.

This chapter aims at putting the general discussion of web archiving on the research agenda. The focus is on the consequences of the archiving process for the Internet scholar, and not on the archiving process itself. Hence, the point of view is that of the Internet researcher, and the archived web is basically discussed as a medium and a text.[1]

The Internet scholar who intends to make use of archived web material, no matter how it has been created, and in which archive it is found, faces two fundamental questions that will also serve as guides in the chapter: (1) is the archived web document a new kind of document? – and if "yes," then, (2) must this document be treated in new ways when used as an object of study? Both questions revolve around "the new" and whether new media necessarily call for new methods and new theories. I will argue in this chapter that, in general, new media are not necessarily new (different) just because they are new (newcomers). And if they are not different from existing media in any significant way, we do not have to approach them with new methods or seek new explanations by the use of new theories. Therefore, the advent of "new" media only implies that we question, on the one hand, their possible differences from existing media and, on the other

The Handbook of Internet Studies Edited by Mia Consalvo and Charles Ess

hand, our well-known theories in the light of the existence of the new media. We will find, however, that the archiving of the Internet in fact entails new methods. First, the Internet as a type of medium is characterized by being dynamic, ephemeral, changing, etc., in other words – in relation to archiving – it is fundamentally different from any other known media type, and therefore it must be approached in different ways than we are used to when we archive; second, when it is archived, it is also fundamentally different from well-known media types, and again we have to approach it in new ways.

We will see this, however, only following a careful discussion of the two questions above. And for this discussion, it must first be determined what we mean by web archiving. In exploring this matter, I further present a brief history of web archiving.

Web Archiving and Archiving Strategies

In order to avoid limiting the field of web archiving to the understanding that has been predominant for the last 10–12 years, I would suggest the following broad definition of web archiving: Any form of deliberate and purposive preserving of web material. This definition will be elaborated by explaining each of its constituents in more detail.

Micro- and macro-archiving

Regardless of the form, the purpose, and the kind of web material preserved, a distinction may be drawn between micro- and macro-archiving. Micro-archiving means archiving carried out on a small scale by "amateurs" on the basis of an immediate, here-and-now need to preserve an object of study. In contrast, macro-archiving is carried out on a large scale by institutions with professional technical expertise at their disposal, in order to archive, for instance, cultural heritage in general (cf. Brügger, 2005, pp. 10–11; cf. also Masanès, 2006, pp. 213–14). Micro-archiving is, for instance, what an Internet researcher does when she uses her own PC to archive the Internet newspapers to be studied, while macro-archiving is performed by, for instance, national libraries who use a complex setup of computers and software with a view to securing the Internet activity related to a nation-state for the years to come.

Web material

Web material has two general characteristics: It is digital and it is present on the Internet. It can therefore be considered a sub-set of, on the one hand, digital media (e.g. e-print, e-books, computer game consoles, CD-ROM/DVDs, etc.), and, on the other hand, the Internet (an infrastructure with a variety of protocols, software types, etc. – Usenet, Gopher, Internet relay chat, email, etc.). Thus, web material is the specific part of digital media and of the Internet that is related to the use of protocols and markup languages originating from the world wide web in the broadest sense.

However, "web material" is still a very broad term that needs to be clarified further. A framework for doing this is to identify some of the main analytical levels on which we can talk about web material as a delimited signifying unity (cf. Brügger, 2009). At one end of a five-level scale we have all the material that is present on the *web as a whole*, and at the other end of the scale an *individual web element* on a webpage, such as an image. And in between, we can focus on different analytical clusters of web material, of which the most significant are: the *web sphere*, which is "a set of dynamically defined digital resources spanning multiple Web sites deemed relevant or related to a central event, concept, or theme" (Schneider & Foot, 2006, p. 20); the *website*, which is a coherent unity of web pages; and the individual *webpage*. Following this five-level stratification, the preserved web material can thus be anything from the web as a whole to a web sphere, a website, a webpage or a single web element on a webpage, and these five strata are mutually each other's context.

One last distinction has to be made. Normally web material must have been made public on the web. But, in fact, material that does not meet this criterion might also deserve to be archived. At least the following three types should be taken into consideration: (1) non- or semi-public material: i.e. web material kept on an intranet, where it is only accessible to a limited and known number of people; (2) pre-public material: i.e. design outlines, dummies, beta versions, etc.; (3) public material that has been published in other media types, for instance in printed media (newspapers, magazines, books, journals) or television (commercials, TV spots, etc.). Especially with regard to the early period of web history, the only preserved sources are these indirect, non-digital pieces of evidence (this last category is thus the exception to the above-mentioned point that all web material is digital). With this last addition in mind, web material can therefore also be material that has not been made public on the web for one reason or another, or material that stems from the web but that we only can access in other types of media.

Deliberate and purposive preserving

That web archiving is considered a deliberate and purposive act means that one has to be conscious, first, *that* one is preserving the material at all, and, second, *why* the web material is being preserved. The simple act of putting an html-file on a web server that is connected to the Internet in order to publish it is, in fact, a way of preserving the file. Still, it is not done deliberately with a view to archiving the file, but simply as an "unconscious" and integrated part of making the file public (although the result is the same, as long as the file is not removed or the web server is not closed).

And in order to be designated as web archiving, the casual storing of web material is not sufficient. In addition, the act of preserving must be accompanied by some degree of reflection on why the archiving is carried out – be that for fun;

with a view to preserving the old family home page; in order to document how a certain webpage looked at a certain point in time; with a research project in mind; or with a view to preserving the cultural heritage of a nation or a culture.

Forms of web preserving

By far the most widespread way of web preserving is *web harvesting*, which in short means that a web crawler downloads files from a web server by following links from a given starting point URL. Web harvesting has been the predominant approach to web archiving in all major (inter)national archiving projects, just as it has been setting the agenda for most of the literature (e.g. Masanès, 2006; Brown, 2006).

However, two other forms of web preserving have to be mentioned in order to complete the picture, although their use is not yet very widespread, at least not among larger archiving institutions. The first can be distinguished from web harvesting by the fact that it transforms the html text into *images*. One of the simplest ways of preserving web material is to capture what is present on the screen or in a browser window by saving it either as a static image or in the form of moving images by recording a screen movie.[2]

Delivery is another way of getting web material archived. In contrast to web harvesting, where the web material is retrieved from the "outside" by contacting the web server, the material can be delivered from the "inside," that is directly from the producer. From the point of view of the archive, one can distinguish between proactive and reactive delivery: proactive delivery being based on some kind of prior agreement with the producer and made from a certain point in time and onwards, while reactive delivery is random delivery of old material that the producer, a researcher, or the like delivers to an archive without prior agreement.[3]

Each of these three forms of web archiving has advantages as well as disadvantages. In short: web harvesting is very useful with large amounts of web material, and it can to a great extent be automated, though far from completely; indeed, the more automated, the more deficient is the archived material as to missing elements and functions, and some types of elements can only be harvested with great difficulty. In addition it is possible to keep the link structure intact to some extent, and harvested material is easier to use afterwards since it is searchable; it looks much like the live web (but as we shall see this is far from the case, cf. the section below, *The Archived Web Document*), and it can be treated automatically as a digital corpus, which is why it is well suited for use in research projects where these demands have to be met, such as projects involving content analysis, link and network analysis, sociological analysis, or ethnographical analysis.

Screen capture is very useful for smaller amounts of web material. The capture of static images can to a certain extent be automated, whereas the capture of screen movies requires the active presence of an individual. Both types of screen capture have the advantage that all textual elements on the individual webpage are preserved, and screen movies have the further advantage of being capable of preserving some of the web forms that web harvesting has great difficulty in archiving, e.g. streaming media, computer games, user interaction, personalized websites, etc. It is also possible to preserve working links, with the disadvantage, however, for the screen movie that a later user has to follow the movements made by the individual who made the recording. Screen capture is very useful when one wishes to study style and design, carry out rhetorical or textual analysis, or in general carry out studies where access to the html code is not important.

The overall advantage of delivered web material is that it can provide access to web material that has not yet been archived, and makes it possible to complete existing web archives. The disadvantage is that delivered material is often very heterogeneous and it can be difficult to reconstruct it meaningfully, or to integrate it in an already existing archival structure: this is a serious problem with proactive delivery and even more so for reactive delivery.

Archiving strategies

As indicated by this brief presentation of the constituents in the definition of web archiving, a complete web archiving that preserves web material exactly as it was on the live web is very much an exception. To a great extent, archiving involves deciding to what extent we can accept omissions in our archived web document, and this fundamental condition forces us to formulate and choose between different archiving strategies, each focusing on specific web material, purposes, and forms of archiving, and each with their gain and loss. The following three strategies have been used by larger web archives that all base their collections on web harvesting. (For examples of each strategy see below, the section headed *The dynamic web in (trans)national web archives*).[4]

The *snapshot* strategy intends to take a snapshot of a certain portion of the web at a certain point in time. This strategy is often used for the harvest of a large number of websites, where either no specific websites have been selected prior to the archiving, or a very general selection has been made, based on, for instance, domain names (e.g. country codes such as .uk, .fr). In general, archiving using the snapshot strategy is not very deep, and normally it is not accompanied by a subsequent quality control and an eventual supplementary archiving, which is why the result is often of erratic quality – the price to pay for the large number of websites. Although this strategy is called "snapshot," it can take up to several months to archive something like a national domain, which is why it cannot be used for rapidly updated websites.

The *selective* strategy sets out to archive a limited number of websites that have been selected individually prior to archiving, because they are considered important, because they are supposedly updated with short intervals, because one wants to carry out a deep archiving, or a combination of these reasons; it is often accompanied by a quality control and an eventual supplementary archiving, which is why this type of material is in general of good quality, but for a limited number of websites. The selective strategy is normally used for harvests with very short intervals (hours, days), meaning that this strategy has a more continuous form.

The *event* strategy aims at archiving web activity in relation to an event in the broadest sense of the word, for instance, general elections, sports events, catastrophes, etc.; in other words planned as well as unplanned events. It can be considered a combination of the two above-mentioned, since the intention is to archive a larger number of websites than the selective strategy, and with a higher frequency than the snapshot strategy.

A Brief History of Web Archiving

Three main phases can be distinguished in the brief history of web archiving. However, the history of web archiving is so short – 10–12 years – that it can be difficult to draw clear lines, which is why the phases in many cases persist side by side with various degrees of overlaps. In addition there are great national differences.[5]

The pre-history of web archiving

In the very first years of the existence of the web, individuals, families, organizations, and institutions from time to time felt a need to preserve what they had created or what they met on their way through what was once called "cyberspace," mostly preserved as either a number of html-files or screenshots. In this amateur era, archiving was random and based on here-and-now needs, just as the act of preserving was not accompanied by reflections as to what was done in terms of archiving and it was not considered part of a greater, systematic effort with a view to preserving the cultural heritage.

There exists no systematic overview of the existence of this kind of material, but a lot of it is presumably either in the possession of the people who created it, or scattered all over today's web. Famous examples are Tim Berners-Lee's first browser screenshot from 1990 and the least recently modified webpage (1990).[6] These amateur archivings are in many cases the only existing evidence of the early web (together with the above-mentioned non-digital media types), and this early approach to web archiving lives on in the many forms of micro-archiving.

Static web publications in national libraries

In the same period a few of the major national libraries begin to preserve material that has been published on the web. Thus, an increased professionalism as well as a clearly formulated intention of preserving the national digital cultural heritage emerge. However, since these initiatives are carried out by libraries, the overall approach is very much that of print culture. Focus is on the nexus between legal deposit law[7] and a conventional understanding of publishing (delivery by or download from publishing houses), on well-known library infrastructures (cataloguing, organizing), and based on a conception of web material as primarily static documents, like books or periodicals that just happen to be published on the web.

The first example of this approach is the Electronic Publications Pilot Project (EPPP) conducted by the National Library of Canada in 1994 (EPPP, 1996, pp. 3–4). Here the mindset of traditional libraries is applied to the web, at least in this early pilot project: primarily static web material that looks like printed publications is archived (journals, articles, etc.) just as the archived material is catalogued. However, later on, web documents such as websites and blogs are also archived by the National Library of Canada, based on a selective strategy and on submission of the publication from the publisher. The traditional librarian approach with focus on static web publications continues parallel to the third phase, namely different forms of web harvesting.

The dynamic web in (trans)national web archives

As a spin-off of search engine technology, web crawlers were developed, and what proved to be the most important phase in the brief history of web archiving was born. In this phase, the number of archiving initiatives increases dramatically. The professional approach and the ambition of preserving the digital cultural heritage are still predominant, but the libraries are no longer the only archiving institutions, and their manner of conceptualizing the material worthy of being preserved is being challenged. In this period the idea of archiving virtually anything that can be found on the web is formulated, with regard for neither the kind of web document in question nor who has made it public – not only material that looks like printed publications from publishers is to be archived, but also dynamic web material.

However, in terms of strategy, there are substantial differences among the major initiatives as to how this overall aim is pursued. The following four examples illustrate this, each being the first of its kind.

The Internet Archive was created in 1996 as a non-profit organization, located in The Presidio of San Francisco, and with the purpose of preserving historical collections that exist in digital format, among others the web. The web collection is built on web crawls done by the for-profit company Alexa Internet, and crawls are done according to link data and usage trails – what is crawled is where the links point and where the users go, which is why the Internet Archive is transnational by nature (cf. Kimpton & Ubois, 2006, pp. 202–4). The main archiving strategy is the *snapshot* approach, where all the web material that the web crawler encounters is archived every eight weeks. The material is not catalogued in a library sense, but is organized much as it looked on the web. However, the Internet Archive actually started its collection by making an *event*-based collection of the websites of all the 1996 Presidential candidates, but this event harvest is not conceptualized as such; it was done for more strategic reasons, in order to demonstrate the potential value of preserving webpages at a point in time where this value was not self-evident to everyone (cf. Kimpton & Ubois, 2006, p. 202).

The Swedish web archiving project Kulturarw3 was initiated by the Royal Library (national library of Sweden); it was inaugurated in 1996 and made its first harvests in the summer of 1997 (Arvidson, 2001, pp. 101–2). It is also based on a *snapshot* strategy, but in contrast to the Internet Archive, it is the first snapshot-based archive that delimits the archive by the borders of a nation state, since the main objective is to archive everything Swedish on the web (for a definition of "Swedish," see Arvidson, 2001, p. 101).

The Pandora archive was created in 1996 by the National Library of Australia, and uses the *selective* strategy with a view to preserving "significant Australian web sites and web-based on-line publications" (Cathro et al., 2001, p. 107). On the one hand, a limited number of websites are selected, archived, and catalogued, in which Pandora resembles the first library-inspired collections (like the Canadian EPPP project). But on the other hand, the archive is similar to the two above-mentioned inasmuch as entire websites are also archived from the very beginning (from 2005 the National Library of Australia has also used the snapshot strategy).

The Danish Internet archive Netarchive.dk is a joint venture between the State and University Library and The Royal Library, the two Danish national libraries, and was created in 2005, when a new legal deposit law came into force (from 1997, only static web publications were to be submitted to the national libraries). Although Netarchive.dk was not created until almost 10 years after the above mentioned, it was nevertheless the first web archive to formulate an overall strategy combining all three strategies, which means making four annual snapshots of the Danish portion of the web, selective harvests of 80 websites (on a daily basis), and two or three annual event harvests (Jacobsen, 2007, pp. 1–4).

With the Internet Archive as the exception with its transnational scope, each of these first national initiatives has been taken up by several other countries. In addition, several small-scale special archives and collections have been created (some of which are organized in national consortiums), in the same way as archiving on a commercial basis and aimed at both companies and scholars has appeared.[8]

Concurrently with the growing number of web archives, still more formalized forums and institutionalized professional forms of cooperation emerge, especially among the primary players, i.e., technicians and librarians. In 1994, the Commission on Preservation and Access and the Research Libraries Group created the Task Force on Digital Archiving, which in 1996 completed the report *Preserving Digital Information*. In 1997 the Nordic national libraries founded the working group the Nordic Web Archive, and in 1998 the European initiative Networked European Deposit Library (Nedlib) was created by eight European national libraries, a national archive, ICT organizations, and publishers. In 2003 11 major national libraries and the Internet Archive joined forces in the creation of the International Internet Preservation Consortium, and finally in 2004 the European Archive was created, based on partnerships with libraries, museums, and other collection bodies. In addition, the field of web archiving has for some years now had its annual conferences, email lists, websites, monographs, edited volumes, etc.[9]

Gradually members of the Internet research community that might be expected to use the archived material are beginning to show interest in being more closely involved in the discussions of web archiving. Some of the first examples of collaboration between scholars and web-archiving institutions are the Dutch project Archipol (2000), the conference "Preserving the Present for the Future" (2001) and webarchivist.org (2001).[10]

The Archived Web Document

No matter how an archived web document has been created, and no matter in what archive it is found, the Internet scholar cannot expect it to be an identical copy on a 1:1 scale of what was actually on the live web at a given time. In this respect the archived website differs significantly from other known archived media types. This point can be clarified by focusing on two interrelated clusters of characteristics of the archived web document: on the one hand, it is an *actively created and subjective reconstruction*; on the other hand, it is almost always *deficient*.[11]

An actively created and subjective reconstruction

Archived web material has one fundamental characteristic, regardless of the type of web material in question and regardless of the form and strategy of the archiving process: it is an actively created subjective reconstruction. First, the simple fact that a choice has to be made between different archiving forms and strategies, both in general and in detail, implies that the archived web document is based on a subjective decision by either an individual or an institution. For instance, it has to be decided which form of archiving has to be used, where the archiving should start, how far from the start-URL the crawl is to continue, whether specific file-types are to be included/excluded (e.g. images, sounds, flash), whether material is to be collected from other servers, how the material is to be preserved, both here-and-now and in a long-term perspective, etc.[12] Second, the archived web document is a reconstruction, in the sense that it is re-created on the basis of a variety of web elements that stem from either the live web or the producer, and that are reassembled and recombined in the archive.

Thus, the archived web document is the result of an active process that takes place at the nexus of the "raw material" present on the web, and a number of choices with regard to selecting and recombining the bits and pieces at hand. In this sense the archived web document does not exist prior to the act of archiving; it is only created in a stable form through the archiving process. In this respect the archived web document stands apart from other media types. When archiving newspapers, film, radio, and television, the main choices are related to the selection of the material, while the archiving process itself *grosso modo* consists of taking a copy out of circulation and storing it; no matter who stacks the newspapers or presses the record button on the video recorder, the archived copies are identical to what was once in circulation, just as all copies are identical. In contrast, with web material, choices have to be made in relation to both selecting and archiving, and we always do more than just remove the web material from circulation; the material is never totally unchanged.

Deficiencies

The other cluster of characteristics of the archived web document is that it is almost always deficient when compared to what was on the live web. Apart from the deficiencies caused by deliberate omissions, two other sources of deficiencies can be singled out: those related to time, and those caused by technological problems during the process of archiving.

One of the major reasons for deficiencies in relation to time is what could be called the dynamics of updating, that is the fact that the web content might have changed during the process of archiving, and we do not know if, where, and when this happens (cf. Schneider & Foot, 2004, p. 115; Brügger, 2005, pp. 21–27; Masanès, 2006, pp. 12–16). A brief example can illustrate this:

> During the Olympics in Sydney in 2000, I wanted to save the website of the Danish newspaper *JyllandsPosten*. I began at the first level, the front page, on which I could read that the Danish badminton player Camilla Martin would play in the finals a half hour later. My computer took about an hour to save this first level, after which time I wanted to download the second level, "Olympics 2000." But on the front page of this section, I could already read the result of the badminton finals (she lost). The website was – as a whole – not the same as when I had started; it had changed in the time it took to archive it, and I could now read the result on the front page, where the match was previously only announced. (Brügger, 2005, pp. 22–3)

As this example illustrates, it is obvious that the archived web document is deficient, since it is incomplete compared to what was once on the live web – something is lost in the process of archiving, due to the asynchrony between updating and archiving. But it is also deficient in another and less obvious way: since the archived web document is not only incomplete, it is also "too complete" – something that was not on the live web at the same time, the content on two webpages or website sections, is now combined in the archive, and it is difficult to determine what the website actually looked like at a given point in time on the basis of these two. The consequence is that the archived web document is in danger of being subject to the following paradoxical double inconsistency: "on the one hand, the archive is not exactly as the website *really* was in the past (we have lost something), but on the other, the archive may be exactly as the Internet *never* was in the past (we get something different)" (Brügger, 2005, p. 23). And we have great difficulty in determining which of these two is actually the case, if either. In this fundamental way we cannot rely on archived web documents to be identical to what was once on the live web. On the one hand, this problem is minimized the smaller the web material in question and the longer the intervals between the updating; on the other, the problem remains latent, and even worse, we do not know if there is in fact a problem or not.

If the archived web document has been created on the basis of web material delivered from the producer instead of harvested material, we are confronted with other problems related to time, and thereby to other kinds of deficiencies. Delivered material is often complex and fragmented, and it can be almost impossible to reconstruct a meaningful and datable entity out of, for instance, a collection of graphic files, the database of a Content Management System, or the like. The temporally based deficiencies in relation to delivered material are not caused by an asynchrony between two archived elements from different points in time, but rather by the impossibility of assigning more fragments to one fixed point in time at all (cf. Brügger, 2008a, p. 156).

Apart from the possible deficiencies caused by the dynamics of updating or the possible fragmented character of delivered material, the archived web document is also very likely to be deficient due to more technical reasons (software or hardware). For instance, words, images/graphics, sounds, moving images can be missing, or some of the possibilities of interaction can be non-functional in the archived web document.[13]

In summary, it can be argued that since the actual act of preserving web material in almost all cases changes the material that was on the web in a number of ways, the process of archiving creates a unique *version* and not a copy. Two consequences follow from this.

The first is the very obvious but often neglected fact that if several archived versions of, for instance, a given website exist from the same date they will probably differ from one another. Although this consequence might seem obvious, it has not so far attracted much attention in the literature on web archiving. However, a test conducted as part of the research project “The History of www.dr.dk, 1996–2006” clearly shows that versions of the same website that had been archived on the same date by different archives showed great differences in all respects (cf. Brügger, 2008b).

The second consequence is that the archived web document is a version of an original that we can never expect to find in the form it actually took on the live web; neither can we find an original among the different versions, nor can we reconstruct an original based on the different versions.

Web Philology and the Use of Archived Web Material

As it has been shown, the archived web document is a new kind of document, since it differs significantly from other types of archived documents in a number of ways. Taking these new conditions into account, must the web scholar then treat the archived web document in new ways compared to well-known media types and material on the live web? As shall be argued below, the answer to this question is "yes."[14]

Web philology

Since we are always dealing with a version, the major problem is to determine how close each version is to what was actually on the live web at a specific point in time (day, hour). Still, we must be aware that this can only be determined with various degrees of probability, not with certainty. One of the best ways of increasing the probability is to compare several existing versions that are as close as possible to each other in terms of time of archiving; but in doing this we can only base our comparisons on a study of the differences and similarities between existing versions, and not between an original and the versions. In this task we are in many ways close to the textual criticism of the philology of manuscripts (manuscript books as well as draft manuscripts) where a variant is also compared to a variant, without any authoritative original at hand. Some of the basic approaches of classical textual philology are probably of relevance to what could be called web philology; however, they must be brought into line with the media material specificity of the archived web document.

In short, the task of the manuscript scholar is to examine, on the one hand, the differences and similarities between a variety of ancient manuscripts with a view to determining whether one or more of them constitutes a source text, and on the other hand to clarify the relations between variants succeeding in time in the past, i.e. their provenance and affiliation backwards in time. (See Figure 2.1. The arrows indicate the direction of examination.)

Since the Internet scholar takes up another type of document based on a different media materiality than written or printed text on parchment or paper, she is confronted with other kinds of problems. First, she cannot set out to examine the differences and similarities between existing versions with a view to establishing

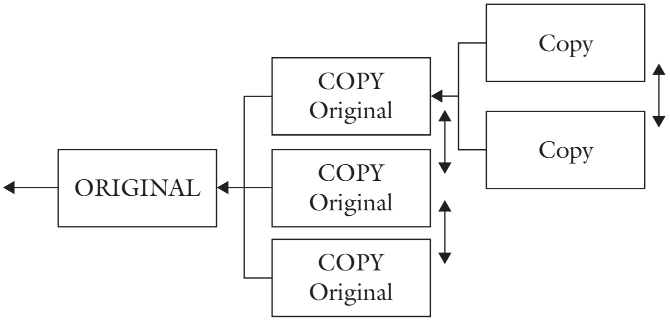

Figure 2.1 The Examination of Manuscripts

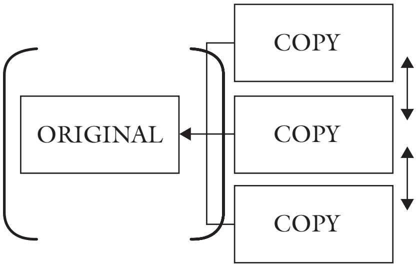

Figure 2.2 The Examination of Archived Websites

one of them as a source text, since none of them is likely to be identical to what was once on the live web. Second, she is examining versions that might have been created at almost the same point of time, which is why she has to trace differences and similarities in simultaneity rather than backwards in time. Third, and as a consequence of the first two points, she can only take one step back in time, insofar as she can be said to examine things backwards in time at all. (See Figure 2.2.)

Apart from these three points, the archived web document has a number of constraints and possibilities that make its examination different from that of the written manuscript, due to the media materiality of digital writing (cf. Finnemann, 1999, pp. 142–8; Brügger, 2002, p. 21). First, the archived web document is a multi-leveled text, in the sense that it can be examined on levels ranging from the immediately perceptible text (the signifying units that we see/hear) to the variety of underlying textual levels that are not immediately perceptible (the source code (HTML, XML, etc.), as well as the layers of the Internet (the TCP/IP model, the OSI model or the like). Second, digital writing makes it possible that the archived web document might have been continuously rewritten in another sense than we know it from manuscripts, since the continuous rewriting can take place after the act of archiving, for instance in relation to long-term preservation, migration to other data formats, etc. Third, digital writing makes it possible to automatically compare archived web documents, at least to some extent. And fourth, exactly identical versions of the same web material can actually exist in different archives (most likely small, non-complicated websites that are rarely updated).

Rules and recommendations

Thus, the web scholar has to make comparisons that differ substantially from those of the classical philologist, and she must therefore proceed in other ways when she sets out to make a consistent statement about how close a given version is to what was on the live web. The following methods and rules might guide the web scholar in this task.[15]

First of all, anyone who uses an archived website must be critical as to his sources. He must be conscious that the version used is not necessarily the only existing one, and that he should therefore try to trace other versions in order to be on more solid ground when determining how close a version is to what was once on the live web.

He also must be prepared for the fact that due to the faulty and deficient character of the archived web document it often has to be navigated differently and examined more closely and by other means than normal on the live web.[16]

And finally, when evaluating the probabilities of what a given web document actually looked like on the live web by comparing different versions, the nature of these comparisons can follow a set of basic rules and procedures. The following six points constitute an attempt to formulate such a set of rules in more detail.[17]

1 The least deficient version as original: one way of getting as close to having an "original" for the comparisons as possible is to use the least deficient version as original. By "original" is not meant "as it was on the web" but just the most complete of the available versions. This "original" could be termed original*.[18]
2 Proximity in time and space: the closer the versions compared are to each other in terms of time (from same version to same day to earlier/later day) and space (same version, same archive, another archive) the greater the possibility of rendering probable what a given textual element actually looked like on the live web.
3 Speed of change: the more stable an element is, the greater the increase in the possibility of rendering probable what it actually looked like on the live web.[19]
4 Types of texts: as we move from the global paratexts of the website via the regional and local paratexts to the text itself, we also move from stable to presumably very frequently changing textual elements, thus decreasing the possibility of rendering probable what the element actually looked like on the live web.[20]
5 Genre characteristics: a given textual element is less likely to have been changed if it is found on a website or sub-site that is supposedly relatively stable in terms of genre.
6 Characteristics typical of the period: it can be relevant to involve knowledge of typical websites from the period of time in question.

These rules provide a set of indicators as to what a given textual element looked like on the live web. Moreover, it can be maintained that the indicators interact by forming a constellation of indicators. One might say that the probability is relatively high that an element actually appeared on the live web as it is in the archive if, for instance, the same element is present at the same position in several versions from the same day in the same archive, and if it is a supposedly stable type of text on a supposedly stable sub-site genre, and if this is supported by knowledge of typical websites from the period. However, in many cases we will supposedly be confronted with constellations of indicators that are neither clearly high nor clearly low in terms of probability, but constitute more conflicting intermediate forms where the indicators point in different directions, thus blurring the picture. In conclusion the following general rule can be formulated: the more indicators that point towards a high probability, the higher the probability becomes in general, and vice versa (cf. Brügger, 2008a, pp. 165–6).

The Future of Web Archiving

As the brief history of web archiving shows, Internet researchers are much better off today than they were five or ten years ago in terms of obtaining access to archived web material. Extensive national and international web archives have been established, and the number of archives, as well as the international knowledge infrastructure (working groups, conferences, email lists, etc.) in the field is rapidly increasing. However, there are still challenges to be dealt with from the point of view of the Internet researcher. A research agenda for the next few years could focus on two main areas: The interplay between web archives and Internet researchers, and the Internet researchers' community itself.

Web archives and Internet researchers

In general, it is important to expand the cooperation between web-archiving institutions and Internet research communities (or to establish such cooperation where it does not yet exist). The most compelling argument for this is that the processes of research and archiving are more closely connected than is the case with other types of media, since the research questions that a researcher intends to examine must to a certain degree be anticipated at the time of archiving. With collaboration, the web researcher has a better chance of getting a useful object of study, while the web archives get users who can actually use what has been archived.

Collaborations can take place in a number of ways, of which the following three seem most obvious. First, collaborations can be an integrated part of the day-to-day operations of the web archive, for instance by associating an advisory board of Internet researchers and other users with the archive. Second, collaborations could be occasional, in relation either to specific research projects or to event archivings, for instance by researchers cooperating with the archiving institutions already in the planning of a research project and not after the funding has been found or the project is halfway through. Third, these two types of collaborations could take place on a global scale in relation to transnational research projects, for instance in relation to events with a (nearly) global impact: wars, catastrophes, sports events such as the Olympic Games, etc.[21]

As regards the interplay between web archives and Internet researchers, three other future challenges should be mentioned. First, a targeted tracing and preservation of the heterogeneous web material that has not been made public on the web should be initiated, possibly in collaboration with other memory institutions such as museums. Second, the web archiving institutions should start to experiment more systematically with the two other forms of web archiving mentioned above, i.e. screen capture and delivery. Both tasks have a view to preserving as much as possible of the cultural heritage of relevance for the history of the web; and just to get an overview of these presumably important areas would be a huge step forward.[22] Third, discussions should be initiated regarding the extent to which the analytical software that is used on the live web can be applied to archived web material, with the specific composition of this material in mind.

The Internet researchers' community

In general, efforts should be made to increase Internet researchers' awareness of the problems related to web archiving. This includes, among others, an increased methodological consciousness – by whom, why, and how has an archived web document been created? – as well as an increased awareness of source criticism, since we are basically dealing with versions.

Methodological developments are also needed, partly in relation to micro-archiving, partly as regards web philology, a task that could be carried out in collaboration with archiving institutions as well as with philologists.

Furthermore, it is important to initiate discussions within the research communities about the specific legal and research-ethical issues related to the use of *archived* web documents, in contrast with Internet material in general. A start would be to map the specific problems of archiving, selecting, access, and subsequent use of the material, as well as the national differences (legal deposit laws, archiving based on prior agreement, crawling ethics, acceptance of robots.txt, copyright, privacy, open/restricted access, etc.).

Since the number of archived web sources in web archives is now gradually increasing, the time might seem right to encourage more Internet researchers to begin making historical studies of the web. As maintained at the beginning of this chapter, historical research is the basis for the understanding of the Internet of yesterday, today, and tomorrow.

And finally, the Internet researchers' community has to confront the challenge of diffusing knowledge of web archiving, both to other disciplines within the social sciences and the humanities (political science, sociology, linguistics, literature, arts, media studies, history, etc.) and to memory institutions in general (museums, archives, etc.). The argument for this is that for some years now the Internet has been an integral part of the communicational infrastructure of our societies. This means that there are phenomena that cannot be analyzed and explained exhaustively if the Internet is not part of the analysis, for instance political movements (the extreme right, non-governmental organizations like Attac, international terrorism), youth culture, artworks, and all kinds of existing media (newspapers, film, radio, and television). In as far as these phenomena are entangled in the Internet in various degrees, an in-depth explanation of these involves the Internet, and therefore the Internet has to be archived. Thus, Internet researchers should make an effort to enter into a dialogue with other research communities about their common interest in preserving the Internet of the past, the present, and the future.

Notes

1 Most of the existing literature brings the process of archiving into focus in either a technical or a librarian sense (questions about what hardware and software to use and how, selection, cataloguing, organizing, etc.), e.g. Masanès (2006) and Brown (2006).

2 For a short discussion of archiving by image capture, see Brügger (2005), pp. 47–53.

3 Masanès talks about server-side archiving (2006), p. 27. However, he discusses neither the delivery of old material nor reactive delivery.

4 For strategies in relation to micro-archiving, see Brügger (2005), pp. 33–60, and in relation to delivery see Brügger (2008b).

5 A comprehensive and systematic history of web archiving that does more than just mention the various archives still needs to be written; the following is only a rough sketch.

6 Cf. http://www.w3.org/History/1994/WWW/Journals/CACM/screensnap2_24c.gif, http://www.w3.org/History/19921103-hypertext/hypertext/WWW/Link.html.

7 Legal deposit law is a law which states that a person who publishes a work is obliged to deliver a copy of it to a deposit institution (normally national libraries). The first legal deposit laws came into force in the seventeenth century, which is why "work" was understood as printed work on paper or the like (book, newspaper, poster, etc). Later, several countries have stretched the law out to include audiovisual media (radio, television), and now the Internet.

8 For an overview of the many different archiving initiatives, see Brown (2006), pp. 8–18; Masanès (2006), pp. 40–45. An updated overview can be found at www.nla.gov.au/padi.

9 Examples are the International Web Archiving Workshop (IWAW, since 2001), the email list web-archive (http://listes.cru.fr/wws/info/web-archive, since 2002), the website Preserving Access to Digital Information (PADI, www.nla.gov.au/padi, since 1996), and the books Brügger (2005), Brown (2006), and Masanès (2006).

10 Archipol is a collaboration between historians from the Documentation Centre for Dutch Political Parties (DNPP) and the University Library Groningen (cf. Hollander & Voerman and www.archipol.nl). The international conference "Preserving the Present for the Future – Strategies for the Internet" (Copenhagen 2001) brought together a variety of members of the user community and persons with technical and library knowledge (cf. Preserving the Present for the Future 2001). Webarchivist.org is a collaborative project at the University of Washington and the SUNY Institute of Technology where researchers since 2001 have worked together with archiving institutions (cf. Schneider 2004 and webarchivist.org).

11 This section presents a short version of the insights in Brügger (2005) and Brügger (2008a).

12 Cf. Brügger (2005), pp. 15–19, 30–31, 61–2; cf. also Schneider & Foot (2004), p. 115; Masanès (2006), pp. 17–18, 76.

13 Cf. the test of the quality of archived web documents in different archives in Brügger (2008b); cf. also the discussion of quality and completeness in Masanès (2006), pp. 38–40, and of the technical problems related to reading and integrating delivered material into an existing archive in Andersen (2007).

14 The considerations on web philology in this section are an abridged version of Brügger (2008a).

15 The methods and rules outlined are produced as part of a work in progress, and have therefore to be tested and discussed further (cf. Brügger 2008a, 2008b). It should

also be stressed that they might not all be relevant in relation to all types of archived web material.

16 A deficient link structure often forces us to navigate, for instance by using a sitemap, by making detours, by clicking on something, etc.; and missing textual elements or functions compel us to make use of the visual marks in the archived document showing that something is missing (mouse-over etc.) or to use the source code to reveal what should have been displayed, etc., cf. Brügger (2008a), pp. 161–2.

17 The six rules are quoted almost *verbatim* from Brügger (2008a), pp. 163–5 where each of them is also elaborated.

18 We owe this convention to Grodzinsky, Miller, & Wolf (2008).

19 This rule is based upon a distinction between stability and high-frequency changes; by "stable" is meant that the same textual element is present at the same position on more than one of the webpages of a website.

20 This rule is based on the assumption that a website is composed of different types of textual elements, respectively texts and paratexts, the latter being the small pieces of text that serve as thresholds to the text itself (menu item, headline, footer, "bread crumbs," etc.) and that make the website coherent on either a local, regional, or global scale. Cf. Brügger (2007a), pp. 75, 84–5 for a definition of "textual element" and pp. 86–7 for a brief discussion of paratexts in relation to the website.

21 There are isolated examples of these three types of collaboration. The Danish web archive Netarchive.dk has an advisory board. Webarchivist.org has collaborated with archiving institutions in relation to both research projects and event harvests. And finally, "The Internet and Elections" was a transnational research initiative (cf. http://oase.uci.kun.nl/~jankow/elections).

22 There are examples of initiatives in these directions. The Danish Web Museum – a national web museum that builds curated exhibitions of websites – has based parts of its collections on web material from other media types (print media, TV spots), just as some of the first web-designer companies have donated material (the museum is hosted by the Danish Museum of Art and Design). And the research project "The History of www.dr.dk, 1996–2006" has traced valuable unpublished as well as older published web material at the producer (both kinds of material have been delivered to an existing archive, see Andersen 2007, Brügger 2008b).

References

Andersen, B. (2007). Integration of non-harvested web data into an existing web archive. Retrieved November 2007 from http://netarkivet.dk/publikationer/IntegrationOfDeliveredData.pdf.

Arvidson, A. (2001). Kulturarw3. In *Preserving the Present for the Future. Conference on Strategies for the Internet. Proceedings.* Copenhagen: Danish National Library Authority. Retrieved February 2008 from http://www.deflink.dk/arkiv/dokumenter2.asp?id=695.

Brown, A. (2006). *Archiving Websites. A Practical Guide for Information Management Professionals.* London: Facet Publishing.

Brügger, N. (2002). Does the materiality of the Internet matter? In N. Brügger & H. Bødker (eds.), *The Internet and Society? Questioning Answers and Answering Questions* (pp. 13–22). Papers from The Centre for Internet Research no. 5. Aarhus: Centre for Internet Research.

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

Brügger, N. (2007). The website as unit of analysis? Bolter and Manovich revisited. In A. Fetveit & G. B. Stald (eds.), *Digital Aesthetics and Communication* (pp. 75–88). Northern Lights: Film and Media Studies Yearbook, vol. 5. Bristol: Intellect.

Brügger, N. (2008a). The archived website and website philology – a new type of historical document? *Nordicom Review*, 29(2), 151–71.

Brügger, N. (2008b). *Archived Websites Between Copies And Versions: Test of Versions In Existing Web Archives.* Papers from the Centre for Internet Research. Aarhus: Centre for Internet Research.

Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society*, 11(1–2), 115–32.

Cathro, W., Webb, C., & Whiting, J. (2001). Archiving the web: The PANDORA archive at the National Library of Australia. In *Preserving the Present for the Future. Conference on strategies for the Internet. Proceedings.* Copenhagen: Danish National Library Authority. Retrieved February 2008 from http://www.deflink.dk/arkiv/dokumenter2.asp?id=695.

EPPP (1996). *Summary of the Final Report.* National Library of Canada, May. Retrieved February 2008 from http://epe.lac-bac.gc.ca/100/200/301/nlc-bnc/eppp_summary-e/ereport.htm.

Finnemann, N. O. (1999). Modernity modernised: The cultural impact of computerisation. In P. A. Mayer (ed.), *Computer, Media and Communication* (pp. 141–59). Oxford: Oxford University Press.

Grodzinsky, F., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology*, 10, 115–21.

Hollander, F. den, & Voerman, G. (eds.) (2003). *Het Web Gevangen. Het Archiveren van de Websites van de Nederlandse Politieke Partijen.* Groningen: Universiteitsbibliotheek Groningen/Documentatiecentrum Nederlandse Politieke Partijen.

Jacobsen, G. (2007). Harvesting the Danish Internet: The first two years. May. Retrieved March 2008 from http://netarkivet.dk/publikationer/CollectingTheDanishInternet_2007.pdf.

Kimpton, M., & Ubois, J. (2006). Year-by-year: From an archive of the Internet to an archive on the Internet. In J. Masanès (ed.), *Web Archiving* (pp. 201–12). Berlin: Springer.

Masanès, J. (ed.) (2006). *Web Archiving.* Berlin: Springer.

Preserving the Present for the Future. Conference on strategies for the Internet. Proceedings (2001). Copenhagen: Danish National Library Authority, 2001. Retrieved February 2008 from http://www.deflink.dk/arkiv/dokumenter2.asp?id=695.

Schneider, S. M. (2004). *Library of Congress. Election 2002 Web Archive Project. Final Project Report.* Research Foundation of the State University of New York/SUNY Institute of Technology/WebArchivist.org. Retrieved February 2008 from http://www.webarchivist.org/911-final-report.pdf.

Schneider, S. M., & Foot, K. A. (2004). The web as an object of study. *New Media & Society*, 6(1), 114–22.

Schneider, S. M., & Foot, K. A. (2006). *Web Campaigning.* Cambridge, MA: MIT Press.