DE GRUYTER
OPEN

10.1515/nor-2017-0183

Nordicom Review 29 (2008) 2, pp. 155-175

The Archived Website and Website Philology

A New Type of Historical Document?

Niels Brügger

Abstract

Website history can be considered an emerging discipline at the intersection between media history and Internet history. In this discipline, the individual website is regarded as the unifying entity of the historical analysis rather than the Internet or the Web. Writing the history of a website involves using many sources and methods similar to those used in writing the history of any other media type. But one document type requires special attention: the *archived website*. This is so because the problems involved in finding, collecting and preserving the website are different from those characterizing the archiving of other types of traces of human activity, including other media types. The primary problem is that the actual act of finding, collecting and preserving changes the website that was on the live web in a number of ways, thus creating a *unique version* of it and not simply a copy. The present article sets out, first, to discuss to what extent the archived website can be considered a new type of historical document and how its characteristics affect the task of the website historian who must later use it; second, the article discusses and attempts to formulate some methodological principles, rules and recommendations for a future critical textual philology of the website.

Keywords: internet, web, website, media history, archiving, archive, philology

Introduction

Website history can be considered an emerging discipline at the intersection between media history and Internet history. In this discipline, the individual website is regarded as the unifying entity of the historical analysis rather than the Internet or the Web.

Website history is a natural part of existing media history; the website can partly be seen as the latest acquisition in the history of co-evolving types of media, from print and analogue media to a whole range of digital media, and website history can partly focus on the same analytical domains as existing media histories have done throughout the years, such as institutional analysis, reception studies, studies of media culture and text analysis in a broad sense (analysis of individual programmes, genres, flow, range of programmes, etc.)[1] However, to some extent the traditional conceptual frameworks need to be re-evaluated.[2]

Website history is also an integral part of both the Internet and Web history. Internet history can be considered the broadest category – covering the history of, for instance, email, usenet, mailing lists, IRC etc. – while web history is a sub-discipline within Internet history dealing with the history of the Web as such, the bloggosphere and so on.[3] Website history, for its part, is a sub-discipline of Web history, and its subjects coincide at least in part with those of both Internet and Web history, as the history of a given website is closely related to the history of both the Internet and the Web. Nevertheless, the history of a website cannot be understood exhaustively solely by using concepts and insights from Internet or Web history (cf. also Brügger 2007a).[4]

Writing the history of a website involves using many sources and methods similar to those used in writing the history of any other media type. For instance, the research project 'The History of www.dr.dk, 1996-2006' (cf. below) will be based on such sources as memos, minutes of meetings, reports, policy documents, correspondences, organizational charts, job descriptions, dummies, retrospective research interviews, biographies, diaries, memorabilia, legal texts and statistics.

But one document type requires special attention: the *archived website*. This is so because the problems involved in finding, collecting and preserving the website are different from those characterizing the archiving of other types of traces of human activity, including other media types.

The primary problem is that the actual act of finding, collecting and preserving *changes* the website that was on the live web in a number of ways, thus creating a *unique version* of it and not simply a copy. Therefore, it could be argued that the archived website constitutes a type of historical document, which in many ways differs significantly from other well-known document types.

The present article sets out, first, to discuss to what extent the archived website can be considered a new type of historical document and how its characteristics affect the task of the website historian who must later use it; second, the article discusses and attempts to formulate some methodological principles, rules and recommendations for a future critical textual philology of the website. Satisfactory answers to these two clusters of methodological questions are necessary stepping-stones on the path towards writing website history.[5]

Finding, Collecting and Preserving Websites

If we maintain that the act of finding, collecting and preserving changes the website that was actually on the live web, it becomes relevant to identify the reasons for this assertion.[6] In overall terms, one can distinguish between six ways of integrating web material in a web archive, and these six ways may be split into two main groups depending on whether the material is archived from the Net or delivered from the producer (some of the sub-groups may overlap).

1. Archiving from the Net:
   a. Harvesting, capturing or filming
   b. Harvesting, capturing or filming 'ghost websites'
   c. Harvesting, capturing or filming websites archived on the Net by the producer
2. Delivery:
   a. Delivery of non-archived material from the producer
   b. Delivery of archived material from the producer
   c. Delivery from other archives[7]

A Subjective Re-construction

Any kind of website archiving, whether done by means of harvesting, capturing and filming or delivery, is a *re-construction* based on the bits and pieces that stem from either the Web or the producer of the website. And this re-construction involves a number of subjective choices and unpredictable coincidences with regard to software, strategy and purpose, integration in an archive and so on (cf. Brügger 2005: 15-19, 30-31, 61-62; cf. also Schneider & Foot 2004: 115; Masanès 2006: 17-18, 76).

For instance, it must be decided whether one intends to harvest, capture or film the website, or employ a combination of these three methods, and if so, which combination? And does one intend to archive the entire website? How many levels should the archiving cover? Are photos, sound and moving images to be included? Is material to be collected from other servers? One must also decide the size of what could be called 'the archival element' (cf. Brügger 2005: 40) and the archiving direction; in other words, how do we intend to 'slice up' the website? By using a variety of start URLs on more than one level? And in what order are the archivings from the start URLs to begin? And, finally, a number of choices must often be made with respect to the subsequent montage of the archived material.

In this way, the archiving of a website stands apart from the archiving of other types of objects, be they physical objects, including types of media such as the printed media, photographs and film, or electronic 'objects', from radio and TV broadcasts to cellular telephone services. Regardless of who stacks yesterday's newspapers at the library, they look the same, just as the TV programme looks the same regardless of who puts in the tape and presses the record button. The subjective element almost exclusively lies in the act of selecting rather than in the archiving process itself, which by and large consists of taking the objects out of circulation and preserving them as they were, unchanged (cf. Brügger 2005: 15-19).[8]

In contrast, archiving a website is an active process from the very beginning, where to some extent the archived website does not exist prior to the act of archiving; it is only created through the archiving process, on the basis of 'raw material' from the Internet or the producer (elsewhere I have used the expression 'a document of the Internet' to refer to the archived (harvested) website, cf. Brügger 2005: 30). Two overall and interrelated consequences follow from this. First, the processes of research – for instance, the writing of website history – and archiving are closely connected, more than we know to be the case for other types of media. Second, archiving a website should be accompanied by deliberations as to method: Why and how has the archived version been created? (cf. Brügger 2005: 31-32, and Schneider & Foot 2004: 115).

The fact that the archived website is a subjective re-construction constitutes a fundamental condition of all of the six distinct ways of integrating material in a website archive. But besides this, each of them is characterized by specific problems.

Archiving from the Net

The problems related to website harvesting have set the agenda for discussions of web archiving for the past ten years, and by now many of the problems have been detailed and to a certain degree solved (for reasons of space, only the problems related to harvesting are in focus in the following pages; problems related to capturing and filming are discussed in Brügger 2005: 47-60).[9] However, the dynamic character of the Internet will constantly be a source of uncertainty in relation to harvesting, above all, the dynamic that I have called 'the dynamic of updating' (Brügger 2005: 21-27; cf. also Schneider & Foot 2004: 115). The dynamic of updating refers to the fact that the content of a website may change during the process of archiving, and we do not know if, where, and when the updates occur, which has the following two consequences:

> The first and rather obvious consequence is that we cannot be sure that we have everything in our archive. We will always have lost something in the asynchronous relationship between updating and archiving. The second consequence is less obvious, but no less serious. Not only do we lose something that was there, we are also in danger of getting something that in a way was *never* there – something that is different from what was really there. My archived version of a newspaper's website can be *a combination* of elements from two (or more) versions that were there at different times – but they were never there at the same time as they might now be in my archive. We thus face the following paradox: on the one hand, the archive is not exactly as the website *really* was in the past (we have lost something), but on the other, the archive may be exactly as the Internet *never* was in the past (we get something different). (Brügger 2005: 23).[10]

A variant of this problem is the fact that the greater the time span between two harvests of a website, the more difficult it is to date all elements on the website to a precise point in time. Thus, we cannot know to what extent the harvesting time is identical to the time of publication; in other words, we do not know how long the website, in all its dimensions, has been the way it is when the harvesting takes place. 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). This characterizes all kinds of cross-sectional harvests made at considerable intervals.

The fact that evidence of human activities in the past cannot be dated to a precise point in time, or that the evidence might have been modified within a certain period of time, is in itself not new. Very often it is only possible to date utensils, documents and the like to a given period of time – an antique Greek vase was used between 200 and 100 BC; a document was written between the 8th and 9th centuries – but within the limits of this period of time, the object is identical to itself and is characterized by a one-dimensional temporal logic: It is the same vase or the same document; we are just unable to say precisely when it was created or used within the period. And even if the object has been subject to changes – the vase may have been re-decorated, the medieval hand-written manuscript overwritten, or pages added – the changes will almost always have left some datable traces, allowing us to date them within the period, based on a one-dimensional temporal logic with a clear 'before' and 'after'.

The situation is different for a website harvested at considerable intervals of time or for the first time. In this case, we have an end date (the time of harvesting) and perhaps a kind of start date (an earlier version), but within this period of time a continuous process of publication may have taken place, and it is very difficult to say if, where, and when this occurred, as in most cases publication does not leave any datable traces. Therefore, such an archived website may not be identical to itself within the period that can be delimited, and it is based on a multi-dimensional temporal logic: some elements on the archived website are from one point in time, other elements from another point in time. Imagine a preserved newspaper composed of bits and pieces from several newspapers from the period prior to the copy in question, but without any temporal indication on the fragments. Concerning the website, we might even encounter a temporal logic that not only extends back into the past, but also forward into the present or the future – for instance, when an older website contains commands that are set to get the current news headlines, weather forecast or the like from a web server, and will therefore get the news from the date of harvesting, and later on perhaps even present the news from the date the archive is consulted – like a year-old newspaper in the library suddenly showing today's weather forecast.

If we look at the second way of integrating web material in an archive, the harvesting of ghost websites, the problem is not the harvesting itself, or rather: the problems are not greater than or different from those encountered in harvesting in general. The problem is partly finding the material, partly dating it. First, we have to become aware of the existence of the material – here we often have to rely on other types of sources such as press releases, other media and other sub-sites. In this task, the producer will not always be much help, because systematic policies for discarding old material are rare. Second, we have to date a website that is harvested for the first time, and, as shown above, it may be difficult to determine how long it has looked the way it does at the time of harvesting.

The problems related to the third method, harvesting websites archived on the Net by the producer, resemble those characterizing the ghost websites; however, the problem of finding the material will probably be less significant, as it has been saved intentionally, which is why there will normally be links to it, just as it will often be easy to identify as archived material if it is marked with watermarks or the like.

Delivery

While harvesting has been setting the agenda in discussions of web archiving for some years, the problems that might be related to the delivery of old web material are unknown quantities, at least to my knowledge.[11] Hence, what is written about delivery in the present article is not based upon thorough discussions and experience; it should rather be considered a provisional catalogue of the possible problems.

In general, the fact that the archive is not in control of the formats of delivery (soft- as well as hardware) will constitute a problem for all three methods of delivery. First, reading the material can be problematic: File types may have gone out of use or it may be impossible to establish an adequate computer environment.[12] Second, integrating delivered material into existing archival structures and procedures can be problematic: must GIF, PDF or QuickTime files be converted to HTML or ARC files? Can an archive from 2007 integrate material from 1996 without automatically dating it to 2007? How can the material be made searchable if it is in a different format than the files of the archive and if it does not provide any kind of text (e.g., no URL, HTML); and must searchable data then be provided manually by the archive? How can material already in the archive be made to interact with delivered web material of the same website from the same date (e.g., individual files, sub-sites)? And so on and so forth.[13]

The problems related to the delivery of web material are supposedly not yet very great, which is probably a consequence of the fact that the number of websites archived by individual scholars (micro or on-demand archiving) is still rather limited. But assuming that website history – and website studies in general – will become more widespread, and that as a result the number of archives will grow, we may witness an increased demand for the mutual exchange of archived web material among scholars (micro to micro), on the one hand, and, on the other, retrospective collecting initiatives launched by the great (inter)national archiving institutions with the aim of supplementing and completing the cultural heritage (micro to macro).[14] Hence, we have good reason to put the problems related to the delivery of web material on the agenda.

If an archive is confronted with the first method of delivery – delivery of non-archived material from the producer – it will probably find the complex, fragmented, varied or unpredictable character of the material problematic. First, it may be difficult to re-create a meaningful unity – a web page, a website – out of a pile of heterogeneous bits and pieces.[15] Second, dating the material may constitute a problem. In contrast to harvesting, the problems are even more difficult in this respect because, unlike the time of harvesting, one cannot even be sure of having one fixed point in time. With harvesting, the problem is not determining a period of time but rather the possible lack of simultaneity and self-identity within the period, but the problem with the delivered material discussed here is that it can be difficult to 'insert' a time period in the fragmented continuum of, for instance, a collection of graphic files, a CMS database or the like. Thus, the problem is not (only) dating the material on a time line, but rather re-creating something that can be related to a time line in a meaningful way, as we cannot be sure that there is any consistent division in either the points or periods of time inscribed in the material.

Different kinds of problems can be expected to arise from the other method of delivery: delivery of archived material from the producer. It will probably have a more systematic character, but this may also constitute a complicating factor if, for instance, an entire computer environment has to be established in order to restore a backup version. It will probably be impossible to integrate such an environment in an existing archive, which is why the archive might have to harvest the restored version. The result will be a kind of 'double archiving' where the website is both re-created and harvested, resulting in two archiving phases, each characterized by specific subjective choices. Besides, documentation may be lacking, which can make it difficult to use the material.

The last method of delivery, from other archives, can entail all the problems mentioned above, but they may be aggravated in connection with collections from micro archives, because the variations in, for example, archiving purposes and the use of archiving software are probably greater among 'amateurs' than professionals.[16] Besides, the integration of corpora – material that has been archived as a whole, often to be used for a specific purpose in a research project – can constitute a special problem. In overall terms, a corpus resemble the strategy of event archiving, and in relation to this kind of material, it is important that the archived websites be marked as part of a corpus so that future users know why the material looks the way it does, as it was created with a specific purpose in mind; any supplementary documentation should also be supplied (e.g., the research questions, plans for the archiving process, the archiving log).

Besides all the problems emanating from both archiving the Net and the delivery of old web material, an overall problem for the archive should be mentioned: that of making the many different types of material in each of the six methods interact in a useful and meaningful way in the same archive, from both a technical and a conceptual point of view (cf. the question of the montage of smaller web elements into a meaningful unity is discussed in Brügger 2005: 39-62).

What Can We Expect to Find in a Web Archive?

As has been argued above, the well-known processes of finding, collecting and preserving appear to be different when the object in question is a website, both on a general level and when we look more closely at the six specific ways of integrating web material in a web archive. However, if one wishes to write website history, this is the way things are, and as website historians, we cannot afford to leave the many bits and pieces of web material without trying to make them useful as sources by integrating them into an archive, one way or another.

If the considerations above are taken as a starting point, then what can the historian who intends to write website history expect to find in a website archive? He or she can expect two things. First, that if several archived copies of a given website exist from the same date, they are very likely to be different from one another (to various degrees). This is because of the elements of subjectivity, creation and coincidence, which generally characterize the process of archiving or integrating a website into a website archive, just as the problems related to each of the six specific ways of integrating material into an archive also play a role. Second, the website historian cannot expect to find an original in the form of the website as it actually looked on the Internet at a given time, neither in the sense of finding an original among the different versions nor in the sense of reconstructing an original on the basis of the different versions. In both cases, this is because a version in the archive may either be lacking something (an 'incomplete' original), or it may consist of something that has never been on the Internet at the same time (a kind of 'incorrect' original due to a possible asynchrony between the website on the Net and the archived version) – and we have great difficulty in determining with certainty if and to what extent a given version is either 'incomplete' or 'incorrect'. In addition, some archiving software writes in the archived files (e.g., HTTrack), while other software does not (e.g., Heritrix).

Test of Versions in Existing Web Archives

If this is what can be expected, the question then arises of whether this is really what one finds when confronted with actual existing archives? In the spring of 2007, the research project 'The History of www.dr.dk, 1996-2006' was started.[17] In the context of the project, a pilot project entitled "Method study of the integration of Internet material directly received from producers" was being carried out, and part of this pilot project was a test examining the appearance of a website that had been archived on the same date (and, if possible, at the same time) in different archives, where it had been integrated on the basis of harvesting, capturing, filming or delivery from the producer.[18] Even if only four versions from one date have been tested thus far, the tendency is clear: with one exception, the four versions showed great differences in all respects.

A Critical Textual Philology of the Website

Because we can expect, on the one hand, that if several archived copies of a given website exist from the same date, they are very likely to be different from one another, and on the other hand, that an original in the form of the website as it actually looked on the Internet at a given time cannot be found, knowledge of an original website to which the different versions can be compared is still wanting, and we must therefore make do with *relative comparisons* based on an analysis of the differences and similarities between the existing copies. Therefore, if we intend to say something about how a given website looked at a given time on the Web – and if we want to do so on the basis of one or more archived websites – we cannot determine this with certainty, but only with various degrees of *probability*. Thus, professing to say what is true or false is not an option, instead we must make do with maybe/maybe not. In this task, we are very close to textual criticism of the philology of manuscripts, in the sense of both manuscript books and draft manuscripts. What might begin to emerge could be called a critical textual philology of the website.

Today, the number of website archives is rather limited, and as a consequence the number of different versions of the same website is also limited. But the problems related to a critical textual philology of the website are still a matter of principle, and their practical relevance is expected to grow as more archives are established, professional as well as non-professional. First, an increased number of archives will also increase the number of versions. Second, one can expect two opposing developments to take place: On the one hand, we can expect more – and more skilled – professional Internet archives, which means better qualities of versions, but on the other hand, we can also expect more non-professional archives to be made by website scholars, which means more versions made by amateurs as well as more versions made for special purposes (e.g., corpora made for a specific research project).

What follows is a brief outline of some of the methodological principles, rules, and concepts that a critical textual philology of the website could use to consistently say something about the probability of how a given website looked at a given time on the Web.[19] The considerations are based upon my work with archived websites from various archives – professional as well as non-professional archives – during the past six years, and their applicability and usefulness are being tested as part of the test mentioned above.

A number of the fundamental concepts and questions in 'classical' textual philology are relevant to a philology of the website, but they have to be brought into line with the archived website as a type of document based on a specific media materiality that is different from papyrus, parchment, paper, and printed books.[20] However, it should be stressed that the present article has no intention of giving a thorough presentation and discussion of textual philology in general, its epistemological history, different methods, schools, etc. The following lines are meant to be nothing more than a brief outline of some general lines of demarcation between 'classical' textual philology and website philology.[21]

From Manuscripts to Archived Websites

The objective of textual philology is basically to examine the differences and similarities between variants of textual evidence that resemble one another – fragments as well as entire works – with a view to either finding the most correct variant(s) or reconstructing a lost original, and to comment on these versions so as to give a critical account of their finding or creation and to make them understandable for future readers.

Within the textual philology of 'old' media, the variants being compared can vary in two respects. First, as regards the *media type,* one can compare variants written on papyrus, parchment or paper to each other (e.g., manuscript books or manuscript drafts), variants written on papyrus, parchment or paper to the printed version, and finally one can compare different printed variants. Second, with respect to the *process of (re)creation* of the variants, we can compare variants that are copies of one another (e.g., different copies of a specific manuscript book), one draft to another, a specific copy or a draft to the printed variant, and a printed variant to another printed variant (cf. Fig. 1).

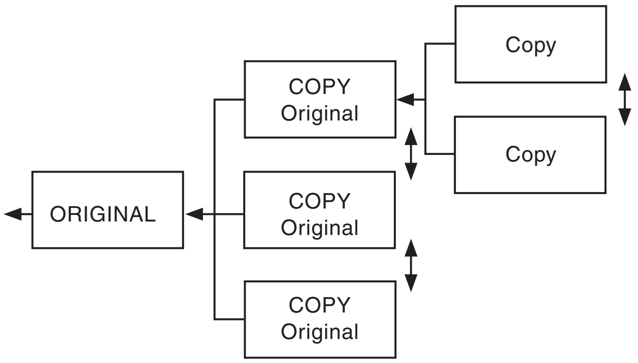
Figure 1. *Comparisons of Variants with Respect to Media Type and the Process of (Re)creation*

	Media type		
Process of (re)creation	manuscript to manuscript	manuscript to printed book	printed book to printed book
copy to copy	+		
draft to draft	+		
copy/draft to printed version		+	
printed version to printed version			+

Figure 1 illustrates that as we move horizontally from manuscript book (written on papyrus, paper, or parchment) to the printed book, and vertically from various copies to the printed version, we move at the same time away from the endless copying and changing towards the stable, authoritative printed version of a book, which constitutes an original in relation to which all earlier copies can be evaluated, and to which all later copies are identical.[22] In this perspective, a textual philology of the website is close to the comparison of a copy to a copy, without any authoritative original at hand, as it is known from manuscript books or manuscript drafts.

The manuscript scholar who today is faced with a number of different ancient manuscript books or drafts knows for certain that they are identical to the manuscripts that actually were created and circulated in the past (given, of course, that they are not forgeries,). He or she 'simply' has to examine: a) the possible differences and similarities between them in order to determine if (and how) one or more of them constitute a source text or an earlier variant of a draft; b) the provenance and affiliation of the different variants backwards in time, because the question of a possible source text is related to variants that succeed one another in the past (cf. Fig. 2).

Figure 2. *The Examination of Manuscripts*

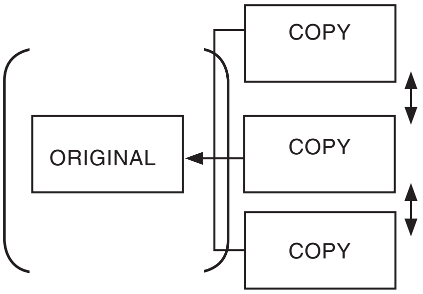


Note: The arrows indicate the direction of examination.

The website philologist, for his/her part, is subject to quite another media materiality, which is why he/she has to deal with other problems. His/her task is different from that of the 'classical' textual philologist in at least the following seven distinct ways (cf. Fig. 3 for the first three points).

1) Because he/she has reason to believe that none of the versions of the archived websites are identical to what was actually on the web in the past, he/she cannot 'just' examine the eventual differences and similarities in order to determine if one/more of them constitute a source text.
2) In contrast to the 'classical' textual philologist who examines versions succeeding in time, the website philologist deals with archived websites that are more likely to be from almost the same point in time (day, hour, etc.), which is why he/she has to trace differences and similarities in simultaneity instead of tracing provenance and affiliation backwards in time.
3) Insofar as the website philologist can be said to examine things backwards in time, he/she only takes one step back, whereas the 'classical' philologist in principle can extend his examination several steps back into the past.

Figure 3. *The Examination of Archived Websites*



Note: The arrows indicate the direction of examination.

4) Due to the materiality of digital writing, the object of the textual philologist of the website – the text – can be examined on several levels. On the one hand, there is the immediately perceptible level, where we see/hear the signifying units directly on the screen or in the speakers. On the other hand, there is the variety of underlying textual levels that are not immediately perceptible, but nevertheless make possible what we see/hear: above all, the source code (HTML, XML, etc.), but also the different layers of the Internet (the TCP/IP model, the OSI model or the like), layers that are all texts written in a digital alphabet with only two letters, 0 and 1 (cf. Finnemann 1999: 142-148; Brügger 2002a: 21). The focus of the present article is mainly on the immediately perceptible layer, but obviously one can advantageously involve the non-perceptible layers.
5) Even if the archived website is preserved in an Internet archive, it may be subject to continuous re-writings in relation to long-term preservation, for instance when it has to be moved to another data format.
6) Due to the digital alphabet, it is possible to some extent to compare archived websites by automatic means (at least on the non-perceptible level).

7) It is possible that two exactly identical versions of the same website can exist in two different archives (most likely small, non-complicated websites that are rarely updated).

Bearing in mind the general considerations mentioned above, we shall now go into more detail about some of the possible methods and rules that can guide the work of the website philologist when he/she needs to say something consistent about the differences and similarities of several versions.[23]

Source Criticism

First of all, the website philologist must keep in mind the possible existence of different archived versions of the same web activity. First, one has to realize that the version one has found is not necessarily the only existing version, second, that one should try to trace other versions (in the same or in other archives), and third, that the more versions that are available, the more likely it is that one can determine how close the different versions are to the website that was once on the live web – and the reverse: the fewer, the more difficult.

Navigating and Examining

An archived website will very often be faulty and defective. However, the faults and deficiencies are not always immediately visible, but are only gradually revealed as one uses the archived website. Thus, use and trouble-shooting coincide, and trouble-shooting is often an ongoing process, which means that very often one has to make use of an archived website in other ways than a website on the live web.

Some of the frequent faults and deficiencies are: first, that the link structure can be broken or defective in other ways, second, that textual elements or functions are missing.[24] In the first case, we have to navigate differently; in the second, we must examine the archived website using unusual means.

Navigating

In the case of a defective link structure, the following phenomena are often encountered, each in their own way forcing us to navigate with inventiveness and care.

a) The link does not work, but the link target is in fact in the archive, which is why what has actually been archived is not always found (this error is often encountered in menu items, which might cause the user to believe that the link target – e.g. a specific sub-site/web page – is not part of the archived version). The following alternative methods of navigating might be used: use a sitemap if possible, make detours (a sub-site/web page is often linked to from more than one web page), click on anything, shift between text and graphic version (in older material), cut off the URL-address from behind (e.g. http://www.dr.dk/Nyheder/Temaer/Oevrige_temaer to http://www.dr.dk/Nyheder/Temaer to http://www.dr.dk/Nyheder).

b) The link is working, but the link target has not been archived, which can have the following consequences: either we are actually informed that the link target is not in the archive (this is the case in the Danish national Internet archive *netarkivet.dk*), or the link takes us to the correct URL but from another point in time, for instance to another day, or to the live web (this is the case with the archive *archive.org*, and the 'distance' between the versions can be up to several years). I use the term 'inconsistent linking' for this last phenomenon, and it is not only due to errors in the archived version itself, but is rather a result of the version being part of an archive with porous boundaries between the versions or to the live Internet. Thus, inconsistent linking adds a new dimension to the unreliability of sources.[25] In the first case, nothing is found because nothing was archived, and there is not much to be done about it; in the second case, you find something that is not part of the version from which you departed, and here we have to navigate with great caution (for instance in the archive *archive.org* it is not indicated very clearly that the link is inconsistent, as only the URL-address in the archive changes, e.g. from http://web.archive.org/web/20000228123340/http://www.yahoo.com to http://web.archive.org/web/20000229135040/http://www.yahoo.com/ (and inconsistent text is not marked).[26]

Examining

An archived website without missing elements or functions is very much the exception. However, valuable information about missing elements is often hidden in the archived website, thus making it possible to extract relevant information despite the fact that the information is not immediately manifest. The following three methods can be used to gain knowledge of missing elements (of the fact that an element is missing as well as what the element was).

a) There is often a visual mark in the archived website, showing that an element – an image, a flash – is missing: Either the name of the file or file type appears, or a mouse-over dialog box communicates the file name or the function that should have been performed.

b) The source code can reveal a great deal about what should have been displayed on the screen or why it looks 'strange' (for instance because a style sheet is missing).

c) When it comes to missing web pages or sub-sites, the site map or the menu items can indicate what might be lacking.

Thus, in a number of ways, the archived website itself communicates its own faults and deficiencies in a somewhat systematic and predictable manner (I use the term 'marked absence' for this phenomenon).

However, we cannot rely on these kinds of marked absences. Sometimes an element is just absent, without any 'explanation' and often in an inconsistent manner (for instance an overall graphical element can be absent on, say, the front page, but present on a web page in a sub-site). Again, the above-mentioned click-on-anything method might come in handy.

All in all, with regard to deficient link structures and missing elements or functions, the overall method is: click on anything, use the source code, and examine every corner of the archived website, even if it appears useless at first glance.

Evaluating Probabilities

As mentioned above, the website philologist can only determine with various degrees of probability what a website or a given textual element actually looked like on the live web – either in terms of the element being present/not present at all, or in terms of the 'content' of the element, that is, *what* is present – and he/she therefore has to make do with relative comparisons. However, the nature of these relative comparisons is to be described in more detail. This will occur, first, by suggesting how relative comparisons can be performed, and second, with a view to guiding and improving these comparisons, by discussing the relevance of involving considerations of the proximity in time and space of the versions, the expected speed of change of the textual elements on the web page, as well as knowledge of the types of texts on the web page, of genre characteristics and of typical websites from the period in question.

The Least Deficient Version as Original

In light of the absence of an original, one way of getting as close to having an 'original' for the comparisons as possible is to use the least deficient version as the original (by 'original' is not meant 'as it was on the web', but just the most complete of the available versions). Thus, one could proceed in the following manner:

1. Evaluate the different available versions with regard to quality: Are all textual elements present, do they work properly, are all sub-sites present, etc.?
2. Choose the most complete as the point of departure for the comparisons.
3. Compare this version to the other versions in order to establish differences and similarities; the comparisons can either be *in toto* (if the similarities dominate) or be focused on specific points (the textual elements, the number of sub-sites, etc.) if the differences dominate, in order to avoid too complex a comparison.

Proximity in Time and Space

When we are to compare several web pages to each other, we could use the following five-step method, where proximity in time and space of the different archived versions is prioritized. We should compare the textual elements on the individual web page to other web pages from:

1. the *same version*
2. another version from the *same day* in the *same archive*
3. another version from the *same day* in *another archive*
4. another version from an *earlier/later day* in the *same archive*
5. another version from an *earlier/later day* in *another archive*

The steps are prioritized from top to bottom, indicating that we lower the probability the more we are forced to move down the list of steps, thus assuming the following rule: The closer the compared versions 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.[27] At first sight, this rule seems to conflict with the requirement mentioned earlier, namely that we should try to trace other versions, and that the more versions we have available, the more likely it is that we can determine how close the different versions are to the website on the live web. This is still true, but the rule of proximity primarily applies when we actually compare the different available versions, if we have been able to locate more than one.

Speed of Change

Based on the assumption that the various textual elements on a web page change at different speeds with regard to both position and content, one can place their expected speed of change within a given period (e.g., a day) on a continuum with two poles: stability and high-frequency changes. If the same textual element is present at the same position on more than one of the web pages of the website (or sub-site), it can be considered a stable element. And if an element is specific to an individual web page, it is more likely to be changing at a high-frequency. These considerations can be a cause for formulating the following rule: The more stable an element is, the greater the increase in the possibility of rendering a probable picture of what it actually looked like on the live web.

Types of Texts

Moreover, a web page can be considered a signifying unit, consisting of a complicated system of texts and paratexts, paratexts being the small textual elements that serve as a threshold to the text itself, either on the same page or on others – a menu item, a headline, interposed excerpts, a footer, 'bread crumbs' that indicate the position of a web page on a website, etc.[28] And these paratexts can form a network on either a local, 'regional', or global scale on a website: For instance, the main menu items that are present on all (or at least the majority) of web pages are global paratexts; some menu items are only present on the web pages of a sub-site, which is why they can be considered regional paratexts, and finally the local paratexts – e.g., specific headlines – are not present on more than one web page (cf. Fig. 4).

Following these lines of thought, we might assume the following rule: 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 high-frequency changing textual elements, thus decreasing the possibility of rendering probable pictures of what the element actually looked like on the live web.

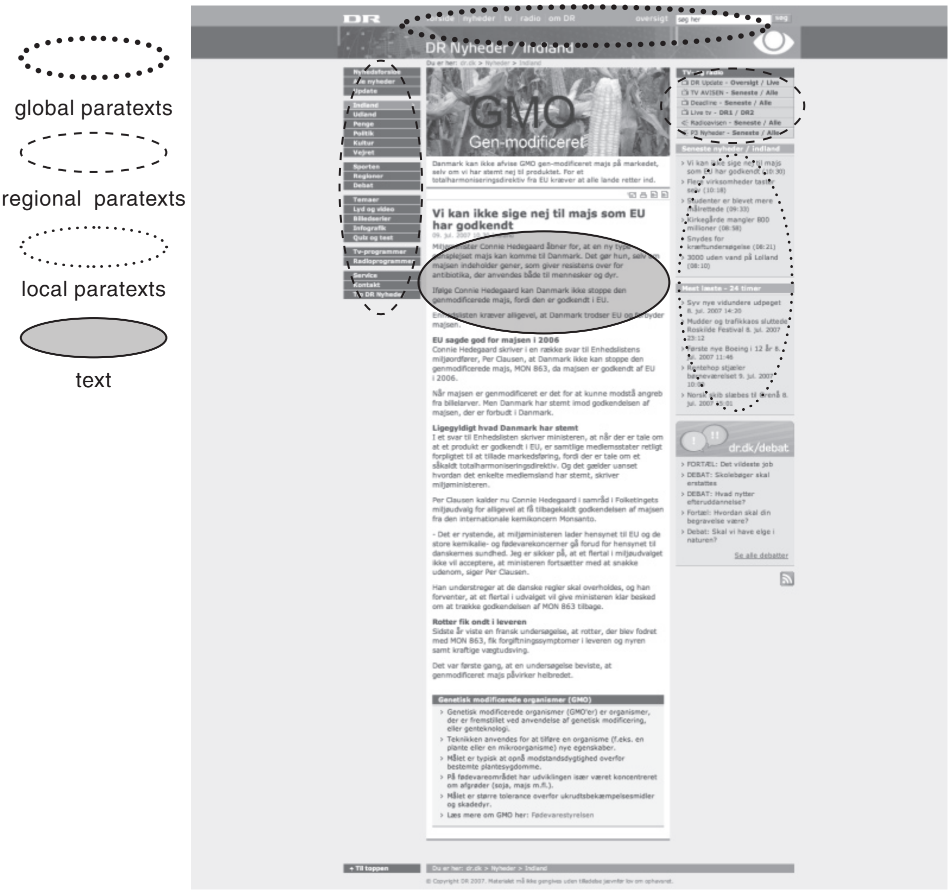
However, concerning the texts themselves (and not the paratexts), different expected speeds of change also seem to be at work, depending on the type of content. A description of a television programme aired a month ago is probably stable; a news item that is continuously replaced or moved further down the web page can be characterized as mezzo stable, and text in debates or chat is supposedly changing at a high frequency.

Finally, one can point to other elements of expression that are presumably relatively stable, such as layout, backgrounds, typography, etc.

Genre Characteristics

The question of expected speed of change also applies to the entire website (or sub-sites) in terms of genre, because a website can as such be positioned on the continuum between stable and high-frequency changes. A few examples can illustrate this. A sub-site with games, etc., for small children is likely to be relatively stable. A sub-site made in relation to a weekly television serial is probably stable when the serial is not being aired, changing once a week when it is being aired. A news sub-site can, on the one hand, change rapidly, but on the other hand it can also have thematic sections (about elections, wars, etc.) that can vary between frequent changes as the event takes place, relative stability as it fades out, and the sub-site can end up being totally unchanged when the theme has ended. And finally, a sub-site with debate is likely to be characterized by high-frequency

Figure 4. *Texts and Paratexts on the Website*



DR forside | nyheder | tv | radio | om DR oversigt søg her søg

DR Nyheder / Indland

Du er her: dr.dk > Nyheder > Indland

Nyhedsforside
Alle nyheder
Update
Indland
Udland
Penge
Politik
Kultur
Vejret
Sporten
Regioner
Debat
Temaer
Lyd og video
Billedserier
Infografik
Quiz og test
Tv-programmer
Radioprogrammer
Service
Kontakt
Til DR Nyheder

Danmark kan ikke afvise GMO gen-modificeret majs på markedet, selv om vi har stemt nej til produktet. For et totalharmoniseringsdirektiv fra EU kræver at alle lande retter ind.

Vi kan ikke sige nej til majs som EU har godkendt

09. jul. 2007 10.30 Indland

Miljøminister Connie Hedegaard åbner for, at en ny type genspleiset majs kan komme til Danmark. Det gør hun, selv om majsen indeholder gener, som giver resistens over for antibiotika, der anvendes både til mennesker og dyr.

Ifølge Connie Hedegaard kan Danmark ikke stoppe den genmodificerede majs, fordi den er godkendt i EU.

Enhedslisten kræver alligevel, at Danmark trodser EU og forbyder majsen.

EU sagde god for majsen i 2006
Connie Hedegaard skriver i en række svar til Enhedslistens miljøordfører, Per Clausen, at Danmark ikke kan stoppe den genmodificerede majs, MON 863, da majsen er godkendt af EU i 2006.

Når majsen er genmodificeret er det for at kunne modstå angreb fra billelarver. Men Danmark har stemt imod godkendelsen af majsen, der er forbudt i Danmark.

Ligegyldigt hvad Danmark har stemt
I et svar til Enhedslisten skriver ministeren, at når der er tale om at et produkt er godkendt i EU, er samtlige medlemsstater retligt forpligtet til at tillade markedsføring, fordi der er tale om et såkaldt totalharmoniseringsdirektiv. Og det gælder uanset hvordan det enkelte medlemsland har stemt, skriver miljøministeren.

Per Clausen kalder nu Connie Hedegaard i samråd i Folketingets miljøudvalg for alligevel at få tilbagekaldt godkendelsen af majsen fra den internationale kemikoncern Monsanto.

- Det er rystende, at miljøministeren lader hensynet til EU og de store kemikalie- og fødevarekoncerner gå forud for hensynet til danskernes sundhed. Jeg er sikker på, at et flertal i miljøudvalget ikke vil acceptere, at ministeren fortsætter med at snakke udenom, siger Per Clausen.

Han understreger at de danske regler skal overholdes, og han forventer, at et flertal i udvalget vil give ministeren klar besked om at trække godkendelsen af MON 863 tilbage.

Rotter fik ondt i leveren
Sidste år viste en fransk undersøgelse, at rotter, der blev fodret med MON 863, fik forgiftningssymptomer i leveren og nyren samt kraftige vægtudsving.

Det var første gang, at en undersøgelse beviste, at genmodificeret majs påvirker helbredet.

Genetisk modificerede organismer (GMO)

- Genetisk modificerede organismer (GMO'er) er organismer, der er fremstillet ved anvendelse af genetisk modificering, eller genteknologi.
- Teknikken anvendes for at tilføre en organisme (f.eks. en plante eller en mikroorganisme) nye egenskaber.
- Målet er typisk at opnå modstandsdygtighed overfor bestemte plantesygdomme.
- På fødevareområdet har udviklingen især været koncentreret om afgrøder (soja, majs m.fl.).
- Målet er større tolerance overfor ukrudtsbekæmpelsesmidler og skadedyr.
- Læs mere om GMO her: Fødevarestyrelsen

TV- og radio
DR Update - Oversigt / Live
TV AVISEN - Seneste / Alle
Deadline - Seneste / Alle
Live tv - DR1 / DR2
Radioavisen - Seneste / Alle
P3 Nyheder - Seneste / Alle

Seneste nyheder / indland
- Vi kan ikke sige nej til majs som EU har godkendt (10:30)
- Flere virksomheder taster selv (10:18)
- Studenter er blevet mere målrettede (09:33)
- Kirkegårde mangler 800 millioner (08:58)
- Snydes for kræftundersøgelse (08:21)
- 3000 uden vand på Lolland (08:10)

Mest læste - 24 timer
- Syv nye vidundere udpeget 8. jul. 2007 14:20
- Mudder og trafikkaos sluttede Roskilde Festival 8. jul. 2007 23:12
- Første nye Boeing i 12 år 8. jul. 2007 11:46
- Rentehop stjæler børneværelset 9. jul. 2007 10:00
- Norsk skib slæbes til Grenå 8. jul. 2007 15:01

dr.dk/debat
- FORTÆL: Det vildeste job
- DEBAT: Skolebøger skal erstattes
- DEBAT: Hvad nytter efteruddannelse?
- Fortæl: Hvordan skal din begravelse være?
- Debat: Skal vi have elge i naturen?

Se alle debatter

+ Til toppen

Du er her: dr.dk > Nyheder > Indland

© Copyright DR 2007. Materialet må ikke gengives uden tilladelse jævnfør lov om ophavsret.

changes. We can therefore formulate the following rule: 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.

Characteristics Typical of the Period

Finally, it can be relevant to involve knowledge of typical websites from the period of time in question. This could involve, for instance, knowledge about the characteristic ways of constructing web pages in general, what menu items, navigation features or other text items may look like and how they may be positioned, use of the paratextual system, etc.

A Constellation of Indicators

When we are to evaluate the probabilities of what a given textual element actually looked like on the live web, we can use the above-mentioned phenomena as a set of indicators. One by one, they can indicate whether the probability is high or low; for instance the probability is relatively high if we can compare web pages from the same version, and correspondingly low if we have to compare web pages from versions from a different day in a different archive, etc. (cf. Fig. 5).

Figure 5. *Indicators of High and Low Probabilities*

Proximity between versions	Distance between versions
Stable textual element	High-frequency changes in textual elements
Stable type of text	High-frequency changes in type of text
Stable sub-site genre	High-frequency changes in sub-site genre
Knowledge of period	No knowledge of period

probability

high ◄——————————► low

Moreover, these indicators also interact. Thus, the probability is relatively high that a given 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. And in between the constellations of indicators that are either clearly high or low in terms of probability, we find a variety of conflicting intermediate forms where the more the indicators point in different directions, the more blurred the picture becomes.

What has been said thus far about the evaluation of probabilities can lead to formulation of the following general rule: the more indicators that point towards a high probability, the higher the probability becomes in general, and vice versa. Thus, using a method that aims at exposing the constellation of the indicators mentioned above enables us to be more precise when evaluating the probabilities of what a given textual element actually looked like on the live web.[29]

Comments and References

In order to facilitate the evaluation of different versions of archived websites, it is recommended that one use the variety of information that in a number of ways 'comments' on the archived website. The following types of information can be of great help.

a) The date and time of the archiving (can be part of the file name or the URL if the archive is accessible on the Internet (e.g., *archive.org*)).

b) Supplementary documents, for instance a plan for the archiving process, a log book or log file (either manually created or automatically generated by the harvesting software); these kinds of documents may contain information about the software used, the starting/ending time and the intervals of the archiving, the starting URLs, the parameters set, errors encountered, whether the archiving is part of a corpus, etc.

c) Other general information, for instance that a given version is part of a selective harvesting that has been performed several times a day.

d) Information about how the material has been integrated into the archive, be it by archiving from the Internet or by delivery from the producer.

The more of these comments that are available, the easier it is for the website scholar to evaluate a version and to explain faults and deficiencies among versions. Besides, this kind of information can turn out to be very important, at least from the scholar's point of view, as existing archives become integrated with one another, for instance when material from the producers or the scholar who has archived a website as an object of study is delivered to a national Internet archive (cf. Brügger 2007a: 10).

This information will also make it easier to make unambiguous references to an archived website. In this regard it is suggested:

- That an archived website be referenced as it has been observed in a given archive, for instance: "www.dr.dk, 30 January 2005, 02:32, as seen in *archive.org*, 23 February 2007, http://web.archive.org/web/20050130023236/http://dr.dk/”
- That the reference always be as precise as possible, preferably referring to a specific web page (in the example above, the web page is the ‘front page’ of *www.dr.dk*)

Conclusion

As a new field, website history – and any kind of history writing that involves an archived website – is still in its infancy, but if it is to become an integrated part of media history as well as of Internet history in general, we must discuss how the archived website can be used as a historical document, and what a website archive should look like if it is to be as useful as possible for media scholars.

To summarize and conclude, it is argued here that any archived website is a reconstruction that does not exist in a stable form before the act of archiving, but is only created through the archiving process on the basis of web elements – both harvested and delivered – that are characterized by a variety of dynamics and temporal logics. Thus, the archived website stands apart from other archived media types, and it challenges the media historian in new ways, with regard to both the creation and the use of archived websites.

The media scholar who sets out to archive a website has to take into consideration, first, that the processes of research and archiving are closely connected, and second, that archiving a website has to be accompanied by methodological deliberations. In other words: What one wishes to examine in a later analysis should to a certain degree already be anticipated at the time of archiving.

And when the media scholar wishes to use archived websites as sources, he/she has to bear in mind that if several archived copies of a given website exist from the same date, they are very likely to be different from one another, and that he/she cannot expect to find an original in the form of the website as it actually looked on the Internet at a given time. Therefore, the archived website has to be treated differently than other well-known types of documents, even – or maybe especially – the website on the live web. The present article has attempted to outline some of the methodological principles, rules and recommendations that can help the media scholar in this task, based on the assumption that a number of the fundamental concepts and questions in ‘classical’ textual philology are relevant to a critical textual philology of the website, and that they have to be brought into line with the specificity of the archived website.

Finally, it should be stressed that the problems of the archived website are not only relevant to website history, as in practice most website studies are based on some kind of archiving of the website with a view to preserving a stable analytical object, except for studies of the live web. Therefore, what has been argued about the special characteristics of the archived website is not limited to website history, but is in fact broader in scope, aimed at website studies in general.

The present work has made only a small contribution to the foundation of website history by putting some of the fundamental methodological issues on the agenda. Obviously, more work is needed in the future to elaborate and refine the theoretical considerations as well as the methodological principles, rules, and recommendations outlined above.

Notes

1. For an overview of some of the theoretical and methodological challenges and discussions within media history, see Bondebjerg 2002; Brügger 2002b; Dahl 2002; Dahl 2004; Djerf-Pierre 2002; Godfrey 2006; Jensen 2002; Salokanges 2002; Schanze 2001; Snickars 2006; Startt & Sloan 1989.
2. For instance, an adequate description of the website cannot be based on well-known text-related analytical concepts such as newspaper article, work, programme etc., just as reception studies must be based on new ideas of interaction. However, these re-evaluations are not only of relevance to website history, but also to website analysis in general.
3. Some of the first major contributors to Internet and web history are Poole 2005; Abbate 2000; Naughton 2002; Gillies & Cailliau 2000; Henderson 2002; Hauben & Hauben 1997.
4. An intermediate analytical level between the Web and the website could be what Schneider & Foot describe as 'the web sphere': "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: 20, cf. also 20-21, 27-35). In web sphere analysis, the focus is less on the individual web site than on the theme that unites various websites (one could maintain that it is a way of conceptualizing web activities that is similar to the archiving strategy of 'event harvesting'; cf. Brügger et al. 2003: 6-7).
5. The underlying and fundamental question of what is understood by 'website' will not be discussed in this article. It should just be mentioned that no matter which criteria one uses to delimit a specific website, historical studies of websites have added a new dimension to 'website studies', because it can be difficult to decide how the website should be delimited over time: Did it have the same boundaries two, five and ten years ago? For a definition of what can be understood by 'website' see Brügger 2007a: 84-87, Brügger 2009 (forthcoming), and Brügger 2008b (forthcoming).
6. For a general introduction to and discussion of web archiving, see Brügger 2008b (forthcoming).
7. These six ways of integrating web material into a web archive are clarified and discussed in more detail in Brügger 2007b: 3-6.
8. Other subjective elements are related to long-term preservation – for instance, when the material has to be moved to another data format or to another media type (e.g., analogue to digital). For a discussion of the problems of long-term preservation of web content, see Masanès 2006: 177-199.
9. Cf. Brown 2006, Brügger et al. 2003, Brügger 2005, Masanès 2006, as well as papers and proceedings from the International Web Archiving Workshop (IWAW, organized since 2001, see www.iwaw.net). Cf. also reports and other activities at the International Internet Preservation Consortium (http://netpreserve.org).
10. The problem of updating is discussed in more detail in relation to the front pages of newspaper websites in Falkenberg 2006: 8-9. What changes is the news item *in toto*, but on both a macro level (logo, overall structure and layout etc.) and a micro level (the actual text of the news item), newspaper websites tend to be rather static, as is evident, for instance, in the fact that the text of the news item is rarely changed once it has been published.
11. This also applies to my book *Archiving Websites. General Considerations and Strategies* (Brügger 2005), which only very briefly refers to delivery-related problems in the first chapter (p. 12). To my knowledge, one of the first experiments with delivery from the producer to an Internet archive was carried out in 2001 in the 'netarkivet.dk' research project, the precursor to the na-

tional Danish Internet archive “netarkivet.dk”, established in 2005 (cf. Brügger et al. 2003: 19, 36). This project has been followed (spring 2007) by another research project: “Method study of the integration of internet material directly received from producers” (cf. the section ‘Test of versions in existing web archives’ in the present article).

12. One of the lessons drawn from the ’netarkivet.dk’ research project (2001) is that it is often necessary to establish an identical copy of the running environment (cf. Brügger et al. 2003: 19).
13. Copyright and economic problems are other problems that could be pointed out (cf. Brügger 2005: 12).
14. Evidence of the former is available on the Air-l-list from May 2007: ”I just started to use Zotero [...]. Now I want to share my grabs with my colleagues. I think they have not yet attacked the sharing of saved pages and annotations. Or do I miss something?” (Posted by Frank Thomas, 10 May). An example of the latter is the retrospective collection of Danish websites that the National Danish Internet Archive ”netarkivet.dk” initiated in 2005, aimed at media scholars.
15. The problems related to the montage of harvested web elements are discussed in relation to micro archiving in Brügger 2005: 39-62. Several of these considerations also apply to the delivery of web material.
16. For instance, a substantial part of my own web archive, begun in 2000, is made with Internet Explorer for Mac (creating WAFF files), which is why this material can only be displayed with this configuration. Besides, the archive also contains a number of screen shots and screen movies (i.e. files that are neither HTML nor ARC).
17. The overall aim of the project is to write the history of the first ten years of the Danish Broadcasting Corporation’s (DR) website, i.e. the period from 1996 to 2006 (dr.dk has been the biggest Danish website for some years). The project is supported by the Danish Research Council for the Humanities in 2007-8. The development of the project website www.drdk.dk is supported by the ”Knowledge Society” research priority area of the Faculty of Humanities, University of Aarhus.
18. The aim of the pilot project is to analyze the archiving and research problems and possibilities connected with the collection of Internet material from the producer and the integration of these in the Danish National Internet Archive ”netarkivet.dk” (”netarkivet.dk” is run by the Royal Library and the State and University library). The pilot project is being carried out in collaboration with the Danish State and University Library/”netarkivet.dk” and is supported by the research foundation of the Ministry of Culture. The technical results of the test with regard to the transformation of non-harvested web data into ARC files are discussed in Andersen 2007. The general results of the test are discussed in Brügger 2008a (forthcoming).
19. As regards the website, the word ’textual’ is meant to refer to coherent units or forms of expression such as written words, still images, moving images, and sound. Therefore, in the following, ‘text’ is understood in a broad sense and is not merely limited to written language. (cf. also Brügger 2007a: 75).
20. Cf. the section ”Finding, collecting and preserving websites” above. For an elaboration of the term ’media materiality’, see Brügger 2002a; Brügger 2002b: 44-52.
21. For a short introduction to and discussion of the different traditions within textual philology, see Cerquiglini 1999: 46-71; for a discussion of some recent tendencies, see the contributions in Tervooren & Wenzel 1997. In addition, it should be noted that when the theoretical and methodological influence between Internet theories and ’classical’ textual philology is on the agenda, focus is mostly on how Internet theories in the broadest sense, for instance hypertext theories, could be used to enhance digital editions of medieval and other manuscripts (e.g., Stolz 2003; Carlquist 2004). The present paper sets out to discuss possible influences in the opposite direction, from ’classical’ textual philology to Internet studies.
22. That the printed copy is called ’authoritative’ is understood in a merely technical way, as all later printed copies of a given edition are alike (cf. also Müller 2006: 187-188). This is a major difference to what characterized the scribe culture of manuscripts, where errors, corruptions and variations were the order of the day. However, errors and corruptions are also a part of print culture, but in contrast to scribe cultures, the number of errors decreases over the years and, most importantly, the printers, editors, and reading communities are conscious of the existence of different copies, a fact of which both the continuous publication of errata and the ’standardization of errors’ (cf. Eisenstein 1996: 51, 59) as well as the actual establishment and development of a critical textual philology are evidence. But although erroneous or variant printed editions can be issued – for instance editions with misspellings, or editions based upon different sources and methodological approaches – once a new printed edition is published, all later copies of it are identical, thus making it ’authoritative’.

23. It should be mentioned that the methods and rules outlined in the following are not all relevant in relation to all types of archived websites.
24. A textual element is understood as a defined coherent textual unit, composed of one of the following four formats of expression: written words (or the like), still images, moving images, or sounds. Textual elements could, for instance, be a headline and the body text (writing), a photograph (image), a banner ad or a news story from television (moving images), or a piece of music (sound) (cf. Brügger 2007b: 84-85).
25. A special type of inconsistent linking is what could be termed 'inconsistent text', namely the fact that a textual element on an archived web page contains commands that are set to get the current news headlines, weather forecast or the like from a web server (cf. the section "Archiving from the net" above). In the archived version, this information will be retrieved automatically from the live web (if the URL still exists), but without making note of the fact. Basically, we are dealing with an inconsistent link, but the content of the link target is seamlessly merged into the text of the archived webpage, which makes the text as such inconsistent, a fact that is not brought to the user's attention.
26. This method of inconsistent linking in *archive.org* is part of a deliberate policy, cf. "Not every date for every site archived is 100% complete. When you are surfing an incomplete archived site the Wayback Machine will grab the closest available date to the one you are in for the links that are missing. In the event that we do not have the link archived at all, the Wayback Machine will look for the link on the live web and grab it if available" (http://www.archive.org/about/faqs.php#202, accessed 12 July 2007).
27. This method presupposes that one has more than one web page from the website in question, which is normally the case if one has access to an archived website. If one has only one web page from a given website (e.g., a pdf file or the like) and neither (parts of) the whole website nor versions from the immediately surrounding days, one is in an unfortunate position.
28. The history of paratexts is discussed by Gérard Genette in Genette 1997, mostly with regard to the printed novel. Danish scholar Finn Frandsen has outlined a development of Genette's insights in order to adapt them to the newspaper, see Frandsen 1991 and 1992. Cf. also Brügger 2007a: 86-87 for a brief discussion of paratexts in relation to the website.
29. This method focuses on the textual elements of the individual web page, but it will probably also be useful in determining the structure of a website, i.e. the presence of sub-sites.

References

Abbate, J. (2000) *Inventing the Internet*. Cambridge, Mass.: The MIT Press.

Andersen, B. (2006) 'Integration of Non-harvested Web Data into an Existing Web Archive', http://netarkivet.dk/publikationer/IntegrationOfDeliveredData.pdf, November 2007.

Bondebjerg, I. (2002) 'Scandinavian Media Histories', *Nordicom Information*, 24, 2-3.

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

Brügger, N. (2002a) 'Does the Materiality of the Internet Matter?', in N. Brügger, H. Bødker (eds.) *The Internet and Society? Questioning Answers and Answering Questions*, papers from The Centre for Internet Research no. 5. Aarhus: The Centre for Internet Research, pp. 13-22.

Brügger, N. (2002b) 'Theoretical Reflections on Media and Media History', in N. Brügger, S. Kolstrup (eds.) *Media History: Theories, Methods, Analysis.* Aarhus: Aarhus University Press.

Brügger, N. et al. (2003) *Experiences and Conclusions from a Pilot Study: Web Archiving of the District and County Elections 2001. Final Report for The Pilot Project 'netarkivet.dk'*, netarkivet.dk, Copenhagen, http://netarkivet.dk/publikationer/webark-final-rapport-2003.pdf (retrieved May 2007)

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

Brügger, N. (2007a) 'The Website as Unit of Analysis? Bolter and Manovich Revisited', in A. Fetveit, G.B. Stald (eds.) 'Digital Aesthetics and Communication', *Northern Lights: Film and Media Studies Yearbook*, vol. 5, Bristol: Intellect, pp. 75-88.

Brügger, N. (2007b) 'Website History: Theoretical and Methodological Problems in an Emerging Field', paper presented at The Association of Internet Researchers Conference Internet Research 8.0: Let's Play, Vancouver, 18-20 October 2007.

Brügger, N. (2008a) (forthcoming) *Archived Websites between Copies and Versions: Test of Versions in Existing Web Archives*. Papers from The Centre for Internet Research: Aarhus: The Centre for Internet Research.

Brügger, N. 2008b (forthcoming) 'Web Archiving – between Past, Present and Future', in R. Burnett, M. Consalvo and C. Ess (eds.) *Blackwell Companion to Internet Studies.* Oxford: Blackwell.

Brügger, N. (2009) (in press) 'Website History and the Website as an Object of Study, *New Media & Society*, 11: 1-2.

Carlquist, J. (2004) 'Medieval Manuscripts, Hypertext and Reading. Visions of Digital Editions', *Literary and Linguistic Computing*, 19, 1, pp. 105-118.

Cerquiglini, B. (1999) *In Praise of the Variant. A Critical History of Philology*. Baltimore/London: The John Hopkins University Press.

Dahl, H.F. (2002) 'The Challenges of Media History', *Nordicom Information*, 24, 2-3.

Dahl, H.F. (2004) *Mediehistorie. Historisk metode i mediefaget*. Oslo: Damm.

Djerf-Pierre, M. (2002) 'The Logic and Practice of Writing Journalism History', *Nordicom Information*, 24, 2-3.

Eisenstein, E.L. (1996) [1993/1983] *The Printing Revolution in Early Modern Europe*. Cambridge: Cambridge University Press.

Falkenberg, V. (2006) 'Metodologiske utfordringer ved arkivering av nettaviser', paper presented at Medieforskerlagets Årskonferanse, Bergen 19-20 October.

Finnemann, N.O. (1999) 'Modernity Modernised – the Cultural Impact of Computerisation', in P.A. Mayer(ed.) *Computer, Media and Communication*. Oxford: Oxford University Press, pp. 141-159

Frandsen, F. (1991) 'Avisens paratekst – et nyt område for medieforskningen', *Mediekultur*, 16, pp. 79-97.

Frandsen, F. (1992) 'News Discourse: The Paratextual Structure of News Texts', in A.-C. Lindeberg, N.E. Enkvist, K. Wikberg (eds.) *Nordic Research on Text and Discourse.* Åbo: Åbo Academy Press, pp. 147-59.

Genette, G. 1997. *Paratexts: Thresholds of Interpretation*, Cambridge: Cambridge University Press.

Gillies, J. & R. Cailliau. (2000) *How the Web was Born. The Story of the World Wide Web*. Oxford: Oxford University Press.

Godfrey, D.G. (2006) *Methods of Historical Analysis in Electronic Media*. Mahwah, NJ: Erlbaum.

Hauben, M. & R. Hauben (1997). *Netizens. On the History and Impact of Usenet and the Internet*. Washington: IEEE Computer Society.

Henderson, H. (2002) *Pioneers of the Internet*. San Diego: Lucent Books.

Jensen, K.B. (2002) 'From Media History to Communication History', *Nordicom Information*, 24, 2-3.

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

Müller, J.-D. (2006) [2002] 'The Body of the Book. The Media Transition from Manuscript to Print', in D. Finkelstein, A. McCleery (eds.) *The Book History Reader*. London/New York: Routledge, pp. 182-189.

Naughton, J. (2002) [1999] *A Brief History of the Future. The Origins of the Internet*. London: Phoenix.

Poole, H.W. (red.). (2005) *The Internet. A Historical Encyclopedia*. Santa Barbara: ABC/Clio.

Salokangas, R. (2002) 'Media History becomes Communication History', *Nordicom Information*, 24, 2-3.

Schanze, H. (2001) *Handbuch der Mediengeschichte*. Stuttgart: Kröner.

Schneider, S.M. & K.A. Foot. (2004) 'The Web as an Object of Study', *New Media & Society*, 6, 1.

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

Snickars, P. (2006) 'Om ny och gammal mediehistoria', *Nordicom Information*, 28, 1.

Startt, J.D & D. Sloan (1989) *Historical Methods in Mass Communication*. Hillsdale: Erlbaum,

Stolz, M. (2003) 'New Philology and New Phylogeny: Aspects of a Critical Electronic Edition of Wolfram's *Parzival*', *Literary and Linguistic Computing*, 18, 2, pp. 105-118.

Tervooren, H. & H. Wenzel (eds.) (1997) *Philologie als Textwissenschaft. Alte und neue Horizonte* (*Zeitschrift für deutsche Philologie*, 116. band, Sonderheft) Berlin/Bielefeld/München: Erich Schmidt Verlag.