# Web History and the Web as a Historical Source

Niels Brügger

*The web and tomorrow's historiography.* Since the 1990s the world wide web (or simply, the web) has been an integral and important part of the communicative infrastructure of modern societies. On the one hand the web has developed as a new medium in its own right, in continuation of other media types such as newspapers, film, radio and television. On the other hand, the web has been intimately entangled in the social, cultural and political life taking place outside of the web. For example, within the realm of politics the web has been essential for the extreme left and right since the mid 1990s (as a platform for discussion and mobilisation as well as for the diffusion of political ideas). And in everyday life an important part of modern youth culture has for a number of years been closely connected to such web phenomena as YouTube, Facebook and Twitter.

Thus, we may expect that in the years to come the web will constitute an important object of study for media and communication historians – as well as historians in general – who want to get an in-depth understanding of important issues in the recent past. With this assumption in mind, the web is very likely to be used as a source for historical studies, and therefore it is important to draw attention to the specific challenges involved in using the web as a historical source.

*Web History.* It is useful to distinguish what web history is from what it is not. First, web history is not internet history. Although many internet users equate 'the web' with 'the internet', the two are not identical. The web is a specific part of the internet, namely the part of the internet using the www-protocol(s), which was invented in the beginning of the 1990s.[1] It is important to bear this distinction in mind since most of the challenges related to using the web as a source are not relevant when talking about the internet in general.

Second, web history is not digital history. Almost simultaneously with the early diffusion of the web, historians imagined that it could be of great importance to historiography. Under the heading 'digital history',[2] the web has often

[1] Cf. the definition in Niels Brügger, Web History: An Emerging Field of Study, in: Brügger (ed.), *Web History*, New York 2010, pp. 1-25, here p. 2.

[2] Cf. Daniel Cohen/Roy Rosenzweig, *Digital History. A Guide to Gathering, Preserving, and Presenting the Past on the Web*, Philadelphia 2006; Andrew McMichael/Roy Rosenzweig/Michael O'Malley, Historians and the Web: A Beginner's Guide, in: *Perspectives* 34 (1996), pp. 11-16.

Zeithistorische Forschungen/Studies in Contemporary History 9 (2012), S. 316-325

ISSN 1612–6033

been considered to have two functions: on the one hand it can help historians in their research processes (searching in digital collections, managing the sources, contacting a forum with fellow historians, etc.), and on the other the web can be used to communicate research results in new ways. Thus, 'digital history' is mainly concerned with the use of the web as a research and dissemination tool, whereas the web as an object of study or as a possible historical source in its own right is not on the agenda. Hence, web history can be understood as historical studies that study the web, either as an object of study or as a source (or both).

*The web.* I suggest that when studying the web – today or in a historical perspective – we should focus our study on five different web strata: the web element, the web page, the website, the web sphere and the web as such.[3] A web element could, for instance, be an image on a web page, a web page is whatever is present in a browser window, a website is a cluster of interrelated web pages, a web sphere is the web activity related to an event, a theme or the like (for instance political elections, catastrophes, etc.) and the web as such is everything that transcends the entire web (for instance the technical infrastructure, organisations such as W3C, etc.).[4]

*The web of the past.* No matter which of the five strata we focus on, the source types that are well known to any historian can be used to substantiate the historical analysis: material objects as well as semiotic sources; and, in relation to the semiotic sources: sources that were handed down to historians as well as sources that are created today. Hence, studies of the history of the web can be based on old computers and internet connections (material objects), written, audio, visual or audio-visual documents from the past (semiotic sources, handed down) or retrospective research interviews (semiotic sources, created today). And just as most of these source types are well known to any historian, so are the methodological challenges related to their use: Where does the source come from? Who created it, and why? How was it handed down to us? How can it be used to answer the research question? Etc.

However, one source type inhabits a special position: the web of the past. The major problem with the web of the past is that contrary to what we might think, yesterday's and today's web is often gone tomorrow if no one has preserved it, for instance a scholar or an archiving institution. Either it has been removed from the web or it has been changed. At the beginning of the millennium it was estimated that the average lifespan of web material was between two and four months.[5] This estimate should be taken with a grain of salt, but

[3] Cf. Brügger, Web History (fn. 1), pp. 3-4.
[4] There are historical studies of each of these web strata, cf. ibid., p. 3.

nevertheless it indicates that although the web can be considered a storage medium of our civilisation, it does not preserve itself for the future – the old web cannot always be found on the web.

It is by no means unusual that material objects, documents and the like disappear – which is why we have institutions such as archives, libraries and museums to collect and preserve them for future research. And any historian is used to making do with incomplete collections of sources. But the specific nature of the web entails, on the one hand, that when it is gone it is gone in such a way that it can be difficult to re-establish it again by the use of the available source types, and, on the other hand, even when it is archived the archiving process itself poses a number of new challenges for the web historian who is going to use it later. Thus, there are good reasons to have a closer look at these two clusters of methodological challenges that are added to the already known ones.

*The missing web.* The main problem for the web historian is not (only) that the web content of the past is missing, but rather that what is also missing is a number of web features which either constitute an integral part of our using and navigating the web, or which provides us with valuable background information about what the web looks like at a given point in time. I shall try to illustrate this point. When studying today's web it is not unusual to add value to the analysis by relating and comparing the part of the web that is being scrutinised – e.g. a web element, a web page, a website, etc. – to the web context into which it is embedded. For instance, if we are studying a specific website and want to evaluate its importance in a web sphere or on the web as such, we can supplement our analysis by: 1) the number of users of this specific website compared to the number of users of other websites, 2) the number of internet users in the country where the website is located, 3) knowledge about the role of the website in a network of websites, for instance in relation to a specific event such as a parliamentary election.

It is not difficult to establish this kind of background for our website analysis since the web itself offers a number of features to help us. The number of users can be found on online services counting user traffic (for instance <http://www.alexa.com> or similar services in different countries); the number of internet users in each country can also be found on the web (for instance at <http://www.internetworldstats.com>); and with a view to knowing the role of a website in a network we can perform a search query (for instance searching for 'parliamentary election'), which can then help us identify the websites to include in the network analysis of in- and outgoing hyperlinks by the use of analytical software.

[5] For a discussion of the different estimates, cf. Marieke Guy, What's the average lifespan of a Web page? 12 August 2009, URL: <http://jiscpowr.jiscinvolve.org/wp/2009/08/12/whats-the-average-lifespan-of-a-web-page>.

But as soon as we want to establish this kind of background in the past, it is almost impossible to do so since the sources mentioned above only exist online, and they only exist here and now. User statistics are usually not kept or made accessible in any book or paper form, and they only provide a picture of things as they look right now (or maybe a couple of months back). And as for the network analysis of hyperlinks, the web historian will miss two things: first, it is impossible to make a search query on the web as it looked five years earlier, and it is thus difficult to identify the websites to include in the network; second, even if one succeeds in establishing which websites to include (for instance by the use of other sources), it is impossible to analyse their network since the entire hyperlink structure is gone.

Therefore, as web historians move backwards in time, they will miss a number of the features and information sources that are taken for granted when using and analysing today's web, and that helps us map the web background in which the entity we are studying is embedded. Since these sources only exist as online web sources, they have in general been lost along with the web they documented. Some of this valuable information could perhaps partly be re-established, and only with great difficulty (for instance user statistics may have been referred to in other media, or they may have been archived unsystematically), whereas other kinds of information can never be re-established (for instance search queries or the hyperlink structure).

*Web archiving.* Although substantial parts of the web have been lost forever, some things are luckily preserved in web archives. Since the mid-1990s a number of international and national web archives have been founded, and easy-to-use web archiving software has been developed, thus enabling scholars to do their own web archiving.[6] Since a number of the challenges related to the use of archived web material as a source concern the very process of web archiving, it is necessary to outline how an online website becomes an archived website.[7] Web archiving does not equate with digitisation. The process of digitisation makes offline material (preserved on paper, tape or the like) available in a digital format, whereas web archiving collects born-digital online material as it appears on the web, which raises a number of different problems. The web can be archived in a variety of ways, but the most widespread is web harvest-

[6] For an overview of the history of web archiving, see Niels Brügger, Web Archiving – Between Past, Present, and Future, in: Mia Consalvo/Charles Ess (eds), *The Handbook of Internet Studies*, Oxford 2011, pp. 24-42, here pp. 29-32.

[7] The website is used here as an example; some of the points below also apply to the other web strata, but not all. The following section refers to Brügger, Web Archiving (fn. 6), pp. 32-34. For an introduction to web archiving, see also Niels Brügger, *Archiving Websites. General Considerations and Strategies*, Aarhus 2005; Adrian Brown, *Archiving Websites. A Practical Guide for Information Management Professionals*, London 2006; Julien Masanès (ed.), *Web Archiving*, Berlin 2006.

ing, i.e. the retrieving of web material from web servers by the use of crawler software that 'crawls' the web based on a list of web addresses (URLs) to archive.[8] In general it is impossible to archive online web content on a scale of 1:1. There are two clusters of reasons for this.

First, the archived website can be considered an actively created and subjective reconstruction of what was once online. It is subjective in the sense that a number of decisions have to be made by whoever is performing the archiving (institution, scholar): which archiving strategy should be used? Where should the archiving start and stop? Should specific file types be included or excluded? Is the crawler software allowed to retrieve files from web servers outside the initial starting points? Etc. And the archived website is a reconstruction in the sense that it has to be assembled by the use of all the archived bits and pieces, first when they are archived, and later when the material has to be displayed for the user of the archive. Thus, it could be argued that the archived website did not exist before it entered the archive, and in this respect it differs significantly from other media types. No matter by whom and where a newspaper is taken out of circulation or the record button on the tape recorder is pressed, the archived material is identical to the original just as all copies are identical. Selection criteria may differ, but once it is decided what to archive the archiving process does not in itself create different versions.

Second, the archived website is almost always deficient. On the one hand a number of technical problems may arise, thus creating deficient copies: for instance images, graphics or possibilities of interaction can be missing. On the other hand the archived website may be deficient because of what I call 'the dynamic of updating', that is, the fact that the website is updated during the process of archiving, and we do not know if, when and where an update is taking place. This problem can be illustrated by an example: during the Olympics in Sydney in 2000, I wanted to save the website of the Danish newspaper *Jyllands-Posten*. I began at the first level, the front page, on which I could read that the Danish badminton player Camilla Martin would play in the finals half an hour later. My computer took about an hour to save this first level, after which time I wanted to download the second level, 'Olympics 2000'. But on the front page of this section, I could already read the result of the badminton finals (she lost). 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 had previously only been announced.[9]

What is in the web archive therefore may prove to be inconsistent with what was once online: obviously, something has been lost, but something may also have been archived which was never online at the same time. In this sense, the

[8] Other archiving methods are screen shots or screen movies.

[9] I already used this example in Brügger, *Archiving Websites* (fn. 7), pp. 22-23.

web archive could turn out to have too little or too much web material, and it can be very hard to determine with certainty what the online web actually looked like at a specific point in time. What we are witnessing here is that even in the age of digital reproduction, we are not only making copies. On the contrary: the web archiving process creates unique versions, each with their individual 'aura'.

*The archived web.* Although the web historian is definitely better off when the web has been archived than when it is missing, even the archived web may cause a number of problems when it is used as a historical source. As suggested above, a web archive is almost always incomplete. In this respect it does not differ from any other archive or collection, since a complete archive is very much an exception – coincidences as to what to include or the deliberate as well as unintended destruction of material have always been the order of the day when making and maintaining archives. However, in many cases a web archive is incomplete in such a way that it is hard to determine if something is missing at all, and if so, what and where. Since these shortcomings are an inherent part of the process of archiving, the archived website mostly does not communicate or document these, and we usually do not have other sources to indicate what might be missing.[10] Thus, the web historian is facing a constitutive uncertainty as to the status of what can be found in a web archive compared to what was once online. Let us now have a closer look at a few of the methodological consequences this entails.

*Web philology and source criticism.* It is important to reconsider the distinction between sources that are handed down to the historian from the past and sources that are created. As I have shown above, the archived website is a reconstruction – it is created, not simply taken out of circulation; but it is not created in the same way as a retrospective research interview about the past, since it is not created independently and without any reference to the online website – on the contrary: it is created by the use of elements which were actually online at a given point in time. With regard to this double character of creation and connectedness to the original online web objects, I have coined the term 'a document of the web'. The archived website is not a document *about* the web – like an interview in which the historian makes conversation with an interviewee about the web of the past – but rather a document *of* the web, that is, a document created by the use of the raw materials (files, etc.) that were actually present on the web at a certain point in time.[11]

[10] However, regarding files the source code sometimes reveals what is missing.
[11] Cf. Brügger, *Archiving Websites* (fn. 7), pp. 30-31.

Hence, the archived website may very well be approached in a manner inspired by the methodological concerns which the retrospective research interview raises. But instead of asking how the interviewee interprets the past or why he or she remembers one thing and forgets another, we have to establish why the archived website actually looks as it does: how did the archiving institution or scholar 'interpret' the online web in the process of archiving? And what was 'forgotten' when archiving it? The ultimate aim of asking these questions is to give as well-argued an answer as possible to the question: what could the website have looked like in the past when it was online? Fortunately – and paradoxically – the web historian can find help in the fact that different versions of the same website may exist (in the same or in different web archives). Since every archived website is a unique version and not an identical copy of what was once online, web historians may be able to clarify what the website looked like by comparing the different archived versions. By doing this their task is in many ways similar to that of the philologist studying handwritten medieval manuscripts with a view to establishing which one is the original (if any).

However, there are a number of differences between the classical philologist and the web philologist, and the most important is that the web philologist's aim is not to find out how one version has been copied in other versions – in the web archive the versions are not copies of each other, but rather copies that are made more or less simultaneously and on the basis of a lost original. Nevertheless, with the characteristics of the archived website in mind we are perhaps in need of supplementing the classical source criticism by a web philology.[12]

*Temporal challenges of the archived website.* The web historian is confronted with another cluster of methodological challenges when using archived web material, namely the problems that originate from the temporal character of the web archive – and its possible temporal inconsistency. Two examples can illustrate this; the first focuses on the website, the other on the web sphere.

One of the most uncomplicated, but often necessary phases of media historians' work is to get an overview over the different archived versions they can study. This is often done by making a register. For instance newspaper historians list the newspapers in a collection, just as a radio or television historian does with the programmes he or she intends to study. However trivial the making of a register may be for media historians, this is a much more complicated matter for web historians since they have to take a number of issues into consideration, one of which relates to the temporal problems of the archived website.[13]

[12] For an outline of a website philology and of some methods and rules to guide it, see Brügger, Web Archiving (fn. 6), pp. 34-38.

When making a register of newspapers or television programmes it is easy to identify the date of publication as well as the start and stop time of a programme. But with a register of websites the start and stop time as well as the interval between them are more complicated to determine in a clear-cut manner. It can be argued that the start time of a website is the point in time when it is published on a web server – but unlike newspapers and programmes, websites usually do not communicate their start time; and identifying the website in a web archive only tells the web historian that it existed at the time it was archived, but not when it was published for the first time.

The time interval which follows the start time is also of a very different nature compared to other types of media, since it usually unfolds as a continuum with no temporal markers indicating when things were published. This implies that in contrast to, for instance, a television station that deliberately marks the time interval of a programme, the subsequent divisions of what happens after the start time of a website have to be inserted either by the web archive or by the web historian, and in both cases randomly and provisionally since they do not originate from the media product itself.

And, finally, when trying to establish the stop time of a website two different criteria can be applied: the website stops either because it is removed from the web server, or because it is no longer updated. However, no matter which criteria are used, it can be hard to determine the stop time of a website when using a web archive. That a website is no longer present in a web archive does not indicate that it no longer exists. The archive may simply have stopped archiving it, and that it is no longer updated is rarely communicated on a website. Thus, the web historian may have to identify the start and the stop time as periods in time instead of points in time.

*Temporal inconsistencies in the archived web sphere.* When the web of the past is missing it is impossible to make a historical web sphere analysis by the use of analytical software, for instance in the form of hyperlink network analysis. But even if the websites we want to include in a network analysis have been archived, the web historian still faces challenges. The problem is that the possible temporal inconsistency of the individual archived website threatens to be repeated and multiplied when moving from the website to the web sphere.

When making a network analysis of the online web, the link structure is mapped as it appears at a specific point in time. But when setting out to make a network analysis in a web archive, the temporal consistency between link

[13] There are a number of other challenges involved in the making of a register of websites; for a detailed discussion, see Niels Brügger, Digital History and a Register of Websites: An Old Practice with New Implications, in: David W. Park/Nicholas W. Jankowski/Steve Jones (eds), *The Long History of New Media. Technology, Historiography, and Contextualizing Newness*, New York 2011, pp. 283-298.

structure and network analysis can by no means be taken for granted. If all entities of the network – usually hyperlinking websites – are not from the same point in time (or close to it), the whole link structure – and thereby the whole network – may have changed from the archiving of the first entity to that of the last one, and in many cases this can be an interval of several days or weeks.

*Conclusion.* The historiographical use of the web as a source is in many ways a challenging enterprise. However, with the assumption in mind that the web is here to stay – at least in the near future – it is important that web historians as well as historians in general begin to debate the fundamental issues related to the use of this new source type.

In my studies of the national Danish public service broadcaster DR's website (<http://www.dr.dk>), archived versions of the website have proven to be an invaluable historical source. On the one hand, it is obvious that when writing the history of a website the website itself is an important source since it is the primary object of study. But, on the other hand, old versions of the website have also turned out to convey important information about the ideas behind the creation and continuous development of the website – ideas that could not be found in the existing printed documents on which the study is also based (minutes of meetings, strategy papers, correspondance, etc.). The study of the history of <http://www.dr.dk> has also shown that if historians want to be sure that today's web is archived, they have to be proactive. DR's website was established in 1996, it has never been archived by the broadcaster and the Danish web archive Netarkivet was not established until 2005 (<http://www.netarkivet.dk>). But I did some archiving of the website myself between 2000 and 2005. In addition, the website was archived irregularly by the American web archive, the Internet Archive (<http://www.archive.org>).

However, in general individual archiving by historians themselves is not always the best solution, since most scholars are not familiar with the ever changing technical challenges involved in web archiving and in the long-term preservation of what has been archived. In the long run these tasks are best handled by professionals, and therefore another way of being proactive is to articulate the need for web archiving with a view to convincing existing cultural heritage institutions and the like to establish local or national web archives.[14] Nevertheless, it is important to stress that technical and librarian skills should be combined with the needs of the scholars who are going to use the web archive with a view to securing that what is archived is also useful for the scholars. For example, in relation to the Danish internet archive Netarkivet, the Ministry of Culture appointed an editorial board with computer scientists,

[14] A number of national web archives have been established during the last decade. For an overview of major existing web archives see <http://netpreserve.org/about/archiveList.php>.

librarians and scholars to advise Netarkivet about their archiving practice. However, scholars' individual needs cannot always be met by the archiving strategies of the web archives. For instance, in many cases only the front page and one level below the front page are archived, which makes it impossible to study an entire website; or websites are not archived often enough, for instance if one wants to study the development of a news website over the course of an entire month. In these cases it is important that some kind of customised on-demand archiving is offered by the web archives.

Thus, not only is the web material in web archives challenging to use, it is also a challenge to get the material archived at all, and in such a way that it can actually be used for historical studies. It is therefore important to bear in mind that web archives do not shape themselves – historians and other scholars are needed to collaborate on this task.

Niels Brügger Ph.D., Centre for Internet Studies, Aarhus University, Helsingforsgade 14, DK-8200 Aarhus N, E-Mail: nb@imv.au.dk