The Web and Digital Humanities: Theoretical and Methodological Concerns

Brügger, Niels; Finnemann, Niels Ole

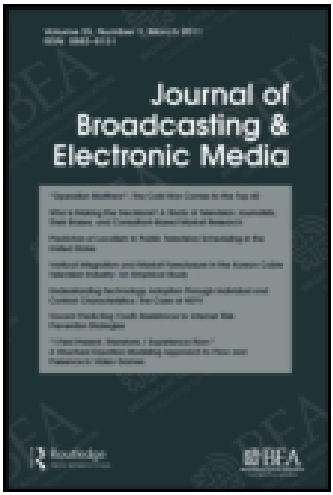
Published in:
Journal of Broadcasting and Electronic Media

DOI:
10.1080/08838151.2012.761699

Publication date:
2013

Document version
Early version, also known as pre-print

Citation for published version (APA):
Brügger, N., & Finnemann, N. O. (2013). The Web and Digital Humanities: Theoretical and Methodological Concerns. *Journal of Broadcasting and Electronic Media*, *57*(1), 66-80.
https://doi.org/10.1080/08838151.2012.761699



Journal of Broadcasting & Electronic Media

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hbem20

The Web and Digital Humanities: Theoretical and Methodological Concerns

Niels Brügger [a] & Niels Ole Finnemann [b]

[a] Internet studies at the Centre for Internet Studies, Aarhus University

[b] Aarhus University

Published online: 12 Mar 2013.

To cite this article: Niels Brügger & Niels Ole Finnemann (2013) The Web and Digital Humanities: Theoretical and Methodological Concerns, Journal of Broadcasting & Electronic Media, 57:1, 66-80, DOI: 10.1080/08838151.2012.761699

To link to this article: http://dx.doi.org/10.1080/08838151.2012.761699

BEA
Educating for Tomorrow's Media

Routledge
Taylor & Francis Group

The Web and Digital Humanities: Theoretical and Methodological Concerns

Niels Brügger and Niels Ole Finnemann

Since the mid-1990s the Web has constituted an increasingly important source for studies of the recent history of society and culture, and a number of national and international Web archiving institutions have been established. This article discusses the different characteristics of Web materials and archived Web materials. It is argued that both of these characteristics differ from the concepts of digital materials developed within the frameworks of digital humanities and that the growing variety of different kinds of digital materials and processes calls for a reinterpretation of the computer, stressing the variability of the functional architecture of digital media.

Today the Internet is the medium which holds the most multifaceted set of materials documenting contemporary social, cultural, and political life. It has become the fulcrum for the general development of media, including mass media and a growing variety of digital devices. If the communicative infrastructure of society in the late 20th century was centered on television, it is today centered on the Internet.

As a variety of digital media penetrates all spheres in society, they also play a still more dominant role for the social sciences, humanities, and arts. They do so in three respects: as archives for contemporary life, as a toolbox for the study of all sorts of digital collections, including digitized collections of non-digital materials (often labeled cultural heritage), and as a means for enforced communication within all spheres of society.

In this article we will concentrate on the first of these aspects, the role of the Internet and particularly its public part as a source material for scholars studying contemporary political, social, and cultural phenomena. The public part of the Internet is mainly found on the Web. The article therefore focuses on Web materials, as

Niels Brügger *(Ph.D., Aarhus University) is a Director and an associate professor of Internet studies at the Centre for Internet Studies, Aarhus University. His research interests include internet theory, Web history, Web archiving, and digital humanities.*

Niels Ole Finnemann *(Dr. Phil., Aarhus University) is a professor of Internet Studies and a Director of NetLab at Aarhus University. His research interests include evolutionary media theory, software supported methods, theories and history of digital media and culture.*

The authors would like to acknowledge the contribution of the COST Action IS1004 www.webdatanet.eu and the COST Action IS0906 ''Transforming Audiences, Transforming Societies.''

DOI: 10.1080/08838151.2012.761699
Journal of Broadcasting & Electronic Media 57(1), 2013, pp. 66–80
ISSN: 0883-8151 print/1550-6878 online

their characteristics will have to inform further reflections on methods and research strategies.

Web materials represent a unique type of source material, and they also include a significant repertoire of links connecting source materials in media otherwise unconnected and unavailable for scholars. Unfortunately the uniqueness of Web materials also includes ongoing changes, modifications, disappearances of content, and reconfigurations of relations. For varying estimates of the average lifespan of Web materials see Guy, 2009.

In so far as Web materials constitute unique source materials, they have to be preserved by some sort of archiving strategy. Otherwise it will not be possible to document what took place on the Internet at any given time. This is a problem for documenting the history of the Internet and the Web and the general history of society. The early years of the Web are already out of sight.

Archiving Web materials is not only of relevance to the study of particular forms of Web communication, but also crucial to other sorts of scholarship and social sciences. Politicians and other (commercial and civic) agencies use the Web to bypass the mass media gatekeepers. Hence, a growing part of public life takes place on a huge network of blogs, debate forums, and semipublic Facebook sites. If these materials are not archived they will be unavailable for future studies of the past. History cannot be written without the documents that allow us to make distinctions between past and present.

For these reasons efforts have been made to collect, store, and preserve Web materials. Some of these initiatives are coordinated by the International Internet Preservation Consortium (IIPC) founded in 2003 (see also Brügger, 2011). Due to the particular nature of Web materials, archived Web materials pose a number of new questions. Web materials differ from most former digital collections, as they are born digital, and archived Web materials differ from both, as they are ''reborn.''

We present three perspectives on digital materials within the humanities in this article. The first two perspectives are well-established within digital humanities, while the third perspective emerges out of Internet studies. From that perspective we suggest a reinterpretation of the notion of the computer and of digital processes due to the growing variety of networked digital media. This is followed by a discussion of the character of Web materials and the character of ''reborn'' archived Web materials.

Digital Humanities and Digital Materials

Within the humanities digital materials and methods have been approached mainly from two different perspectives. One perspective originating in the 1950s and 1960s (e.g., Chomsky) has in recent years developed into what is today labeled ''digital humanities,'' focusing primarily on digitized source materials and ''computational'' methods (the journal *Computers and the Humanities* was founded in 1966. See also Hockey, 2004; McCarty, 2002). In recent years digital humanities has also

included multimodal features and become an umbrella term related to the building of new research infrastructures (see e.g., the report *Research Infrastructures in the Digital Humanities* [European Science Foundation, 2011]). A second perspective develops around the human-computer interaction studies (HCI) and the idea of the computer as a toolbox (Norman & Draper, 1986). While the first perspective originates in relation to mainframe computers, the second perspective originates as an interpretation of the ''personal'' computer in the 1980s. In both cases the standalone computer is taken for granted, and the concept of the computer is more or less explicitly rooted in the idea of digital processes as computational processes. Still, digital humanities is rooted in a framework defined by the digitization of analog sources rather than born-digital materials. Svensson (2011) discusses two US mainstream positions, similar to the two first positions mentioned here, and two extremes: a ''surfer perspective'' and a ''making-code'' perspective.

During the 1990s a third perspective develops around the interpretation of digital media, including a growing array of networked digital devices from the 1990s and onwards.

Within all these perspectives the Web is present as a platform for presentation, distribution, and communication, and they include the use of software-supported and eventually Web-based methods. However, the Web plays no particular role as a source material in its own right in the digital humanities and HCI traditions, while archived Web materials and Web history have only played a minor role in Internet and ''new'' media studies. This is still the case in recent publications such as Berry (2012), Svensson (2011), and Gold (2012). For an exception see Foot & Schneider (2006). Dougherty et al. (2010), and Thomas et al. (2010) discuss scholars' use of Web archives.

From the Idea of Uniform Computers to Digital Media

To bridge these gaps one may want to modify the uniform concept of the computer and ''computing'' to a concept of digital media reflecting the fact that the functional architecture of computers is variable and subject to ever ongoing developments in different directions. Since digital processes are processed in different ways and on different levels of meaning, the notion of computation is also affected.

Such a move seems inevitable to embrace all sorts of digital processes: word, image, and sound processing, all sorts of hypertextual, interactive, and multimodal usages as well as operations in traditional machines, in the growing array of dedicated digital devices, mobile devices, digital circuits in our bodies, in the cities and other surroundings, including processes in our relations to the Internet of things.

What, if any, are the basic and shared characteristics of all these different forms of digital processes?

Our suggestion is that all digital processes share three basic features:

1) Rules as well as data are—at least partly—processed as binary sequences; most digital media are dedicated and not ''universal'' machines, as some parts of

the functional architecture are built into the hardware and not delivered as editable software.
2) The processing is partly controlled by means of algorithms (which are syntactical in nature, specifying a particular functional architecture),
3) The algorithms are controlled on the semantic level of the interface.

In the end the two bits—often labeled 0 and 1,—which are completely void of meaning (similar to letters, rather than numbers) are also the only invariants, while syntax and interface/semantics may be varied endlessly (cf. Finnemann, 1999b).

This definition of digital media deviates, on the one hand, from the concept of the computer as a rule-governed machine which has dominated humanities computing and which is still at work within digital humanities—and rightly so in some particular cases. On the other hand, it deviates from the widespread ''new media'' concept (or the implicit assumption) of the computer as a plastic and freely malleable device that comes with no built-in constraints. The malleability of the computer was introduced in the HCI tradition in the late 1980s (Bolter, 1991; Ehn, 1989) and later became an implicit assumption in the new media studies, where it is hard to find any explicit concept of the computer as a constraining device.

Since the machinery is not conceived of as uniform, the idea of computation as a uniform (mathematical, logical, rule-governed) process is also replaced by the idea of a variable functional architecture. The functional architecture can be modified, suspended, or over-coded on many levels, some of which may be processed on the level of the basic machine architecture, of operating systems, of network connections, of possible configurations of in- and output devices, of programs, of settings, etc. This is compatible with the idea that both ''the digital'' and ''the humanities'' should always come in the plural. To study Web materials scholars from any field will have to use software-supported methods, but scholars from different fields or with different research questions may often want to use them differently and to project different perspectives onto the archived materials.

A main issue here is the particular relation between research questions and the basic criteria for legitimizing the results. In our view digital humanities is at risk to validate itself only due to internal standards. As there is no such thing as ''digital physics'' to be clearly separated from physics, we assume that there is no such thing as ''digital humanities'' to be clearly separated from the variety of humanities disciplines and forms of scholarship, on the one hand, and the variety of software-supported methods developed in studying all sorts of digital sources in all academic fields, on the other. The distinction is questionable both on the side of the ''digital'' and on the side of the ''humanities.''

Web Materials—A Particular Set of Digital Born Materials

Digitized data corpora have been a main source within digital humanities and the antecedent humanities computing traditions. Such materials can be characterized

as homogenous data sources. They are stored in particular formats defined by researchers. The formats chosen may be a general storage format for the types of materials in question (e.g., the Text Encoding Initiative, TEI) or be specifically oriented toward a particular research question. In both cases the basic formats are defined *a posteriori* and extrinsic to the materials, which are usually produced in non-digital form. This is often also the case for many digital born materials such as standardized text and image formats. These characteristics do not apply to Web materials. Contrary to the TEI notion of ''text,'' we use ''text'' in the wide sense, including written, aural, and visual formats. To distinguish between these formats we use ''writing'' or ''word processing,'' referring to the narrow concept of a linguistic text.

Due to the non-proprietary network structure of the Web, allowing everybody to publish software applications as well as any kind of content, eventually personalized by the user or the content provider, Web materials become increasingly heterogeneous in character over the years. The features making digital media and the Internet superior to former media are precisely those that also cause the greatest difficulties for studying and archiving such materials. First, we have the full array of (type)written, visual, and aural formats of non-digital media; next we have their digital equivalents. Digital equivalents differ from their non-digital precursors, because a part of the physical manifestation is replaced by coded sequences, thus changing some formerly physically invariant properties to editable symbolic variables. For words it is not simply the fonts that become replaceable, but also the coded equivalent of the paper. Furthermore, digitized text can be printed in a variety of physical forms. To this comes, however, a number of features and characteristics which are distinct for digital media, features and characteristics which are distinct for Web materials, and finally for a variety of less dominating protocol formats, such as mobile formats like wap, apps, and tweets.

Main Characteristics of Web Materials

Digitized and digital born materials are normally embedded in hypertextual, interactive, and multimodal contexts, defined on the level of the interface and characterizing both standalone machines and networked machines. Digital born materials also include these features in the grammatical repertoire. Hypertext, interactivity, and multimodality are potentially present in all media, but in forms which are particular for each medium. Print media offer the footnote, the table of contents, and other sorts of fixed references which are both similar to and different from digital hypertext. The concepts grew into prominence only in the eras of the PC and of the Internet. Digital hypertext is rooted in the basic searchable address system and random access; interactivity is rooted in the need for input and output formats; multimodality is rooted in the relation between ordinary language and programming language. In the present context of digital media the three notions refer to significant trajectories in the development of digital genres (cf. Finnemann, 2005).

The Web provides a hypertextual, interactive, and multimodal context built into browser interfaces, and it also offers these features as a grammatical repertoire, which may be utilized in a growing variety of forms within any particular site. In the following we delineate the characteristics of Web materials with respect to four dimensions of digital materials: *hypertextuality*, *interactivity*, *multimodality*, and *fluctuation*.

Hypertext was first defined by Ted Nelson in the 1960s as collections of written materials connected by hyperlinks between nodes (Nelson, 1993). Modified concepts played a role in the development of software for standalone PCs in the 1980s (Notecard, Hypercard, Storyspace), closely connected to the development of interactive formats with respect to both content and functional architecture. Not just letters and words, but any unit was now included in the array of hypertextual linking. Some of the hypertextual characteristics mentioned here are discussed in DeWitt & Strasma (1999), Finnemann (1999a), and Kirschenbaum (2000).

The Web protocols widened the reach of hyperlinks to a global scale, and they opened up for a variety of trans-site and in-site applications, such as the organization of a site due to a particular menu structure, and for establishing a multiplicity of routes and passages between elements within any given site and between elements on other sites. The hyperlink is the fundamental glut of the Web, because it connects the visual present with the immense array of hidden materials both within a given site and between sites. The ''spatial'' scaling and reach of hyperlinks is variable, but hypertextual relations may also vary due to the timescale (new links can always be added) and due to authorship relations (existing links and nodes may be modified and new connections and nodes may be added over the years by different authors). Hyperlinks also differ in their semantic nature. They can establish a relation within a work, between works, or between elements in different works. They can be used for associative browsing or goal-oriented navigation. They may be part of a lexical relation, part of the menu structure of a site, or part of fictional relations. Thus, they can be motivated due to different criteria for consistency and overarching ideas. Hyperlink relations may also be motivated as more or less strong author-defined suggestions or as optional choices due to the individual user's motivations, closely related to the variety of possible forms of interactivity. In the end any kind of search due to any kind of search criterion is hypertextual in nature. While a free text search leaves a wide semantic variation open to the user, a meta-tag-based search limits the variation in favor of more goal-oriented link relations. Menu links and navigational links represent a third kind of particular, author-defined utilization of hyperlinks on the Web.

Especially in the 1990s hypertext was seen as a means to break down the organizational hierarchy of printed books, but the use of hypertextual relations on the Web shows that it is rather an indispensable instrument for navigating through increasingly more complex and hierarchically layered amounts of source materials otherwise inaccessible and irretrievable.

The most conventional form of *interactivity* is when a reader is allowed to comment on a text on a given Web page. More distinct for digital media, if compared

to older media, are formats allowing personal, typed communication between individuals, as it is known from chat forums, instant messaging, status updates, tweets, and text messages, whether Web-based or not.

Even more distinct is the array of possible interactions directed toward changes in the functional architecture of the medium. Interactivity has been discussed both from a sociological perspective and in the context of human-computer interaction and computer-mediated communication (cf. Jensen, 2008; Quiring & Schweiger, 2008). Interactivity toward the functional architecture is most often ignored. Still, it is one of the main characteristics of computers as distinct from other mechanical devices.

To prevent unwanted changes a number of restrictions are programmed into the Web sites, but if access is given the source code of the Web page is permanently editable. In principle there is no lower limit for user-generated changes in the functional architecture, in so far as the architecture is defined on the level of the software. However, many devices limit the array of possible changes physically, as they are dedicated to particular purposes by implementing parts of the functional architecture in the hardware. A second limitation is that changes in the functional architecture are risky and beyond the competence of most users—except for the array of possible variations of the settings made available in a menu structure. Still, it can be done by skilled users. It is a significant part of the innovative dynamic on the Web, because the functional architecture may be modified and changed by means of new software on the top of the system. This was exemplified, for instance, by the publication and implementation of the www protocols in the early 1990s (Berners-Lee, 1999; Finnemann, 2005, p. 69).

The variability of the functional architecture also allows the content and service providers to modify the operations of the visitor's machine, to collect information, to deliver personalized services, and to make the services sensitive to the individual messages and actions of any visitor.

However, the usage of these interactive features may vary, both due to the service providers and the users' needs and ideas, but they form a constitutive part of Web materials, thus making these materials distinct from other forms of digital materials.

On a particular Web page the use of hyperlinks and interactive formats can be limited or nearly eliminated. Many Web pages delimit the use of hyperlinks to navigational purposes and refrain from utilizing intrinsic content-sensitive link relations. Still, the hypertextual and interactive features are not simply variable from site to site or from time to time, they also change the very nature of the materials and the set of possible relations between units within the materials.

Multimodal communication is less constitutive for Web communication than hypertext and interactivity. You may do well without it, and word-based communication is by far the most dominating format. Still, multimodal communication may, in a long-term perspective, become as significant as hypertext and interactivity. Most significant is the circulation of still images, graphs, and short videos, predominantly in YouTube format, which also allows for easy embedding on any site. In these cases we still have clearly delimited units of expressions. However, many Web sites also

provide dynamic visualizations, tag clouds, and a growing number of blending in of dynamic features on an otherwise static page.

Fluctuation—A Result of Hypertextual and Interactive Web Dynamics. All sorts of digital materials are in principle permanently editable and reproducible in ways distinct from former media. Units and sequences are delimited by coding, and all units and sequences can be blended deliberately with other units and sequences. In the case of digitized materials one may want to make blends for analytical purposes, but it would be a distortion of the original source and not a legitimate part of the original source material. For Web materials such blends are legitimate parts of the object, such as, for instance, the embedding of a YouTube video in a particular context, or the viral spread of a meme across the Web.

Web materials are permanently editable. They may be remixed, migrated into new contexts and meanings, processes which are often also the offset of the development of new genres, such as the migration of texting into status updates on Facebook and into tweets.

Most of these features apply to all sorts of digital born materials. Distinct for Internet and Web materials are the seamless scales of reach with respect to local-global, public-private (not least many new forms of semi-public spaces), and the array of differentiations of possible connections between sender and receiver. Even if the Web protocols are unidirectional, they are so from the position of both the sender and the receiver.

Finally, Web materials are also distinct compared to digitized materials, as they are born with a particular interface which comes with the materials, even if it is separated and modifiable, but still a part of the materials. Digitized materials are on their side only accessible through interfaces, which are not part of the source, but defined by researchers and archivists as part of the chosen digitization strategy.

A Few Remarks on Methodological Implications

The general characteristics described here may be more or less relevant to particular studies of particular aspects and themes, but they are highly relevant to the understanding of the overall dynamic and complexity of the Web and make it clear that Web materials pose a particular set of methodological issues and document the need to archive Web materials. Some of these issues depend on the particular research question, while others are more generic for a broader range of research perspectives.

First, there is no overarching system of metadata. Many Web sites offer a freely chosen set of keywords, but there are no mechanisms capable of providing consistency, neither in a synchronic nor in a diachronic perspective.

Second, each research project will have to delimit its own timescale and to take into account the stability of the relevant materials.

Third, the identification and delimitation of the relevant materials will often pose problems. Before the breakthrough of the Internet questions such as ''where does

public opinion building take place?'' could be answered by listing the main media and rostra in a society, whereas today it may depend on the particular role of the Web in a particular society, on the issue in question, and on the relevant agencies and their preferred ways of using the Web and particular sites on the Web.

The Challenges of the Archived Web

Web archiving can be performed in a variety of ways regarding archiving purpose and strategy as well as technological choices. On the one end of a continuum we have the very broad Web archiving made by archiving institutions, such as national libraries, aiming at preserving the cultural heritage of, for instance, a nation state. And on the other end one finds the very detailed and narrow archiving in relation to, for instance, a specific research project. In the first case the goal is to archive the Web in such a way that it allows for as many different kinds of research projects as possible in the future, whereas in the latter Web archiving is usually calibrated to fit the research project in question (for an example of the latter, see Foot & Schneider, 2006). In the following we will focus on broad archives such as national Web archives, since future historical studies will to a large extent have to be based on these kinds of already existing Web archives. However, many of the points below also apply to narrow Web archiving.

Compared to digitized as well as born-digital materials, archived Web materials are specific in a variety of ways, each of which raises a number of new questions as to analysis and methods. We do not address issues of copyright and privacy protection, although these issues are very important in relation to Web archiving.

The Characteristics of the Archived Web

The Web Archive is a Real-Time Archive. Since the online Web is changed or deleted at an unprecedented pace, compared to earlier storage media, the selection of what to archive cannot be postponed to future selections. The online Web one wants to preserve must be collected and archived here and now, while it is still online.

The Archived Web is a Reborn, Unique and Deficient Version and Not Simply a Copy of What was Once Online. The archiving institution that wants to archive the online Web must make a number of choices: What to archive and what to omit (specific domain names, types of Web sites, parts of a Web site, archiving depth, specific file types etc.)? Which archiving software and strategy should be used? How to handle updates taking place during the process of archiving? How should the archived material be made accessible in the Web archive? As a consequence of the many possible answers to these questions what is archived is almost never a copy on a 1:1 scale of what was once online; it is rather a collection of unique versions that did not exist before the act of archiving. In addition to the classical

questions regarding selection, which also apply to Web archiving, it has to be taken into consideration that the archived Web material did not exist in the archived form online. It is created in and by the process of archiving, which is why it can be considered ''reborn'' digital material, in contrast to digitized as well as to the born-digital material on which it is based (see also Brügger, 2011, pp. 32–38).

As was the case with digitized data corpora, the archivist decides how the material is integrated in the Web archive and how it can be made accessible for future use. However, one important difference has to be noted: In contrast to digitized material archived Web material always comes with its own interface, namely the interface of the online Web. The Web archive can make its own interface, but it must be on the terms of the Web's original born-digital interface.

The Broad Web Archive is Multitemporal and Multispatial. Since a Web archive usually covers more than one point in past time, numerous versions of the same Web element will exist—a URL, a Web page, an image, a Web site, a hyperlink etc.—each from a different point in time. Thus, in contrast to the online Web where only one copy of each Web element exists at each point in time (even in the case of different user-dependent versions of the same Web page), the time factor implies that the Web archive is multitemporal.

In addition, most often Web sites (especially larger Web sites) are not continuously archived in their totality. At one point in time certain parts are archived; at a later time other parts are archived. In contrast to the online Web where a Web site only has one spatial extension, that is, a certain number of Web pages distributed on the levels below the front page, the spatial extension of the same archived Web site is not necessarily identical throughout time. Thus, the patchwork of Web elements in the Web archive makes it multispatial.

The Web Archive Tends to be Reactive. The rapid and endless new developments of software and use forms on the Web—be that flash, java scripts, Facebook, Twitter, the seamless integration of video on Web pages, etc.—force the archiving institutions to try to keep pace with these changes. However, the scale and the rapid turnover of the changes usually hinder a proactive approach to the fluctuation and dynamics of the Web. Therefore, the Web archive is often reactive in the sense that it is constantly struggling to keep up with the changes. The result is that the Web archive is out of sync for a period of time, until new developments have been discovered and mapped and adequate archiving techniques and routines have been invented, if possible.

Analytical and Methodological Consequences

The above-mentioned four characteristics of the broad Web archive have a number of consequences for the subsequent analytical use of it and for the methods that can be used.

Something is Missing

One of the major consequences is that the Web scholar will probably have to make do with Web material that is incomplete compared to what was once online. That an archive is incomplete is a constitutive element of any kind of archive, since coincidences and deliberate as well as unintended choices often affect what is included in an archive and what is discarded and/or destroyed. However, the question of incompleteness unfolds differently in Web archives compared to digitized material.

Digitized collections can be considered complete in so far as the archive or the scholar selects the material that is to be digitized, and the collection's degree of completeness is usually known when selecting the material. Thus, completeness relates to the analog material itself and not to the fact that it is digitized.

In contrast, the Web archive is characterized by two general types of incompleteness compared to what was once online. First, the user of a Web archive will miss some of the information about the Web which is usually at hand on the online Web, such as search results or information about the current state of the Web (statistics, ranking, number of domain names, users etc.) (cf. Brügger, 2012a). Second, on a more detailed level individual Web elements and possibilities of interaction may be missing in the Web archive, be that streamed audio/video, images, graphics, hyperlinks etc.

The first incompleteness is a result of the fact that the material is no longer online, whereas the latter is often due to an opaque combination of the chosen archiving strategy, deliberate deselections, archiving errors, and technical insufficiencies. What is specific for the incompleteness of Web archives is not that things are missing, but rather that they may be missing in ways which make it very difficult to determine if something is missing at all as well as what and where. First, this is so because, in contrast to many digitized data collections, the Web archive does not have at its disposal a stable original to compare with; the live Web is forever gone. Second, on a detailed level incompleteness is rarely documented—things may simply be missing without explanation. Thus, Web archives as well as the scholars using the archives lack trustworthy methods for determining whether—or to what extent—material in a Web archive is complete.

Heterogeneity and Complexity are Multiplied

The time factor implies that the complex blend of hypertextuality, interactivity, multimediality, and fluctuation which characterizes the born-digital material on the Web at a given point in time increases when combined in the Web archive with material from other points in past time. Thus, each synchronic and historically distinct form of heterogeneity and complexity is multiplied, accumulated, and combined diachronically, and the further the archive stretches back into the past, the more heterogeneous and complex the archived material becomes.

Hyperlinks Become Inconsistent

The structure of hyperlinks is an integrated part of the archived Web and not just an added feature of the archive, such as, for instance, a menu structure or other means of navigation in relation to digitized collections. However, this gives rise to problems of inconsistency related to time and space. First, a temporal inconsistency may exist between the link source and the link target. The link source may have been archived one day and the link target days later. Second, the link target may not have been archived at all, thus giving rise to a spatial inconsistency. In both cases it is difficult to determine if—and to what extent—the archived Web material is inconsistent. And the consequence may be that certain types of studies become very complicated, for instance, historical network analyses of hyperlinks (cf. Brügger, 2012b).

The Archived Web is Edited and Editable

As argued above archived Web material is an edited, ''reborn'' version of what was online. However, once the Web material has entered the Web archive, it is also editable, but in other ways than digitized collections and the online Web, since the archived content itself as well as the division of the material in elements can be changed. Any ''montage'' of the archived elements in the archive—or any extraction from the archive—is also an editing of these elements. This is so because the subdivision of the archived material and the subsequent combination of elements are not necessarily inscribed in the archived material itself as is, for instance, the case with newspaper or television archives where the producer selected the material to be archived (date, hour, section, program, etc.). In contrast, an archived Web site is a continuum with no clear-cut temporal or spatial subdivisions inscribed by the producer; the subdivisions are editable, scalable, and random, and they are made *a posteriori* by either the Web archive or the scholar.

An Archived Web Corpus is a Double Construction

Although a broad Web archive is a collection of Web material, it cannot be considered a corpus, that is, a clearly delimited and structured set of elements (words, texts, images, etc.). When studying the online Web, one can select a corpus to study, for instance, a set of URLs or file types. But since a number of versions of each element exist in a Web archive (cf. above), one has to construct not one corpus, but two. First the URLs that should be included in the study, and second the specific versions of each of these URLs. And because of the limited possibilities of documenting the completeness of a Web archive it can be challenging to construct such a double corpus in an informed way.

Subsequent Processing is Lacking

When making a broad Web archive the traditional subsequent processing made by archivists and librarians (e.g., quality checking or adding metadata) or by scholars (e.g., annotating) is usually not an option—the amount and the complexity of the archived material simply do not allow for systematic and detailed processing of the entire archive once the Web has been archived. Thus, the archive and the scholar have to make do with either the metadata provided by the archived Web itself—for instance, meta-tags in the source code—or with the log files from the archiving process, if the archive makes them available.

The Interoperability Between Web Archives is Challenged

Since the Web is transnational by nature, and since global issues and events transcend the national Web sphere, it would be obvious to undertake transnational and comparative research projects. However, if such studies are to be based on archived Web material, all the analytical and methodological consequences outlined above are multiplied by the number of involved Web archives. The combination and interoperability of existing Web archives may turn out to be one of the major challenges in the future (cf. Brügger, 2012a).

Concluding Discussion

In this article we have mainly focused on the Web, because it has been the main platform for the public use of the Internet so far, but there is a need to consider and eventually include other platforms and forms of digital materials and their possible mutual integration as relevant in the perspective of historical documentation.

With the growing number of more or less dedicated devices, mobile devices, the Internet of things, and an increasing number of online-mediated services, the production of all sorts of information tends to exceed any former limitations. Thus, Web archiving strategies become part of the wider issue of what sorts of information society should keep, how it should be done, and what should be left to oblivion.

We will not enter into the discussion of how to combine the general Web archiving strategies (domain snapshots, different criteria for selective harvesting, and event harvesting), but mention a few game changers of archiving.

First of all digital media and not least the Internet and networked mobile media open up new spaces for stored semiprivate and semipublic communication. Among these forms we find Facebook and Facebook-like forms of social communication, which in some countries play an increasingly influential role in establishing social bonds and public connections. People on Facebook may still be ''bowling alone,'' but they are connected in their everyday life relations (Putnam, 2000). At the same time public opinion building is increasingly influenced by or takes place

on Facebook sites, which thereby become a relevant source for future historians. It is not clear how it would be possible to establish solid archives for these materials, due to the technical, legal, and ethical issues. This is even more so for the relevance of such materials as a source for the study of mental life in our age.

A second game changer is the issue of the delimitation and selection of materials. This is not simply a matter of why and what we want to keep, but also an issue of what it is actually possible to get. A number of limitations are mentioned above. To these should be added the lack of log information and usage patterns, not least the use of hyperlinks. Even if hyperlinks and their functionality are preserved, they do not reveal whether they are used and they reveal even less about the behavioral patterns of the individual users.

While historical perspectives have been maintained within digital humanities as efforts to preserve and analyze cultural heritage, Web materials have been almost completely overlooked, even if they may be considered one of the most significant contemporary contributions to the cultural heritage of mankind. In ''new media'' studies, which have developed outside the Digital humanities and which have in fact been interested in studying Web materials, historical perspectives have been dismissed.

If society still wants to be capable of writing its own history, it will need Web archives and methods for studying these archives. If the humanities should regain its historical position in academia, it will have to be capable of interpreting the stories told in digital materials and among them most significantly archived Web materials, as most other digital materials will vanish in the near future.

References

Berners-Lee, T. (1999). *Weaving the Web*. London, UK: Orion Business Books.

Berry, D. M. (Ed.) (2012). *Understanding digital humanities*. New York, NY: Palgrave Macmillan.

Bolter, J. D. (1991). *Writing space. The computer, hypertext, and the history of writing*. Mahwah, New Jersey: Lawrence Erlbaum.

Brügger, N. (2011). Web archiving—Between past, present, and future. In M. Consalvo, & C. Ess (Eds.), *The handbook of Internet studies* (pp. 24–42). Oxford, UK: Wiley-Blackwell.

Brügger, N. (2012a). Web historiography and Internet studies: Challenges and perspectives. *New Media & Society*, first published on November 21, 2012 as doi: 10.1177/14614448 12462852.

Brügger, N. (2012b). Historical network analysis of the Web. *Social Science Computer Review*, first published on September 6, 2012 as doi: 10.1177/0894439312454267.

DeWitt, S. L. & Strasma, K. (Eds.) (1999). *Contexts, intertexts, and hypertexts*. Cresskill, NJ: Hampton Press.

Dougherty, M., Meyer, E. T., Madsen, C., van den Heuvel, C., Thomas, A. & Wyatt, S. (2010). *Researcher engagement with web archives: State of the art*. London, UK: JISC.G & C.

Ehn P. (1989). *Work-oriented design of computer artifacts*. Stockholm, Sweden: Arbeitslivcentrum.

European Science Foundation (ESF) (2011). *Research infrastructures in the digital humanities*. Strasbourg: European Science Foundation (ESF).

Finnemann, N. O. (1999a). *Hypertext and the representational capacities of the binary alphabet*. Working Paper no. 77-99. Aarhus, Denmark: The Center for Cultural Research.
Finnemann, N. O. (1999b). Modernity modernised. In P. A. Mayer (Ed.), *Computer media and communication—A reader* (pp. 141–160). Oxford, UK: Oxford University Press.
Finnemann, N. O. (2005). The cultural grammar of the Internet. In K. B. Jensen (Ed.), *Interface://Culture—The World Wide Web as political resource and aesthetic form* (pp. 52–71). Copenhagen, Denmark: Samfundslitteratur/Nordicom.
Foot, K. A. & Schneider, S. M. (2006). *Web campaigning*. Cambridge, MA: MIT Press.
Gold, M. K. (Ed.) (2012). *Debates in the digital humanities*. Minneapolis/London: University of Minnesota Press.
Guy, M. (2009). What's the average lifespan of a Web page? *http://jiscpowr.jiscinvolve.org*, August 12. Accessed October 19, 2011. http://jiscpowr.jiscinvolve.org/wp/2009/08/12/whats-the-average-lifespan-of-a-web-page
Hockey, S. (2004). The history of humanities computing. In S. Schreibman, R. Siemens & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 3–19). Oxford, UK: Blackwell.
Jensen, J. F. (2008). The Concept of Interactivity–revisited: Four new typologies for a new media landscape. UXTV '08 *Proceedings of the 1st international conference on Designing interactive user experiences for TV and video*. New York, NY: ACM.
Kirschenbaum, M. G. (2000). Hypertext. In T. Swiss (Ed.), *Unspun: Key concepts for understanding the World Wide Web* (pp. 120–137). New York & London: New York University Press.
McCarty, W. (2002). Humanities Computing: Essential problems, experimental practice. *Literary and Linguistic Computing*, *17*(1), 103–125.
Nelson, T. H. (1993). *Literary machines 93.1*. Sausalito, Ca.: Mindful Press.
Norman, D. & Draper, S. (Eds.) (1986). *User centered system design: New perspectives on human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Putnam, R. D. (2000). *Bowling alone. The collapse and revival of American community*. New York, NY: Simon & Schuster.
Quiring, O. & Schweiger, W. (2008). Interactivity: A review of the concept and a framework for analysis. *Communications*, *33*, 147–167.
Svensson, P. (2011). The digital humanities as a humanities project. *Arts and Humanities in Higher Education*, *11*(1–2), 42–60.
Thomas, A., Meyer, E. T., Dougherty, M., van den Heuvel, C., Madsen, C. & Wyatt, S. (2010). *Researcher engagement with Web archives: Challenges and opportunities for investment*. London, UK: JISC.