Is There a Way Back or Can the Internet Remember its Own History?

Marcus Burkhardt

Abstract
As we shift from analogue to digital media as the predominant means to express ourselves and to communicate with each other, the question how we construct personal and cultural memory in and of cyberspace becomes increasingly important. Considering the ephemeral nature of digital information in the Internet, this paper asks how the vast amounts of digital information in this global communication and information network will be memorized. The paper focuses on the Internet Archive's effort to preserve the entire Internet for future generations. Facing the risk a "Digital Dark Ages", the Internet Archive was founded in 1996 by a group of visionaries around Brewster Kahle, at a time when years of the Webs history already have been lost forever. Converging with the "database logic" of the new media, the Internet Archive does not form a narrative of the Internet's history. Drawing upon a media archaeological approach, some technological and conceptual means underlying the Internet Archive's attempt to preserve the entire Internet is discussed. The paper concludes asking what kind of memory we can gain by accessing the Web Archive.

Key Words: Internet Archive, Digital Heritage, Digital Preservation, Media Archaeology.

*****

1. Introduction

In the 1990s, the Internet quickly became an important part in many of our lives. As we shift from analogue to digital media as the predominant means to express ourselves and to communicate with each other, the question how we construct personal, collective, and cultural memory in and of cyberspace becomes increasingly important. Early in the history of digital media, Ted Nelson put forth a vision of a global information network that he conceived as a never forgetting document space. He began working on his project Xanadu in the 1960ies and it is still not finished today. On the project's website we can read, "We need a way for people to store information not as individual 'files' but as a connected literature. […] Documents must remain accessible indefinitely, safe from any kind of loss […]".[1] Nelson suggested that in the realm of the digital, nothing would need to be lost anymore, as digital media would allow us to create a universal

archive of everything. Nelson's vision entered at least to a certain extent the cultural imaginary of our time. Yet, considering the ephemeral nature of digital information in the Internet, we need to ask how these digital memories will be memorized. Does the Internet have a memory of its own? The paper discusses some chances as well as challenges of reading the digital past in the present. Hereby I will focus on the Internet Archive's efforts "to prevent the Internet […] from disappearing into the past".[2]

2. Digital Heritage

Texts, images, and media products in general cannot be merely understood as documents. They are monuments as well. In 1969, Michel Foucault made this famous claim in his Archaeology of Knowledge.[3] The Foucaultian approach prevailed in German media theory and led to the establishment of a media archaeological line of thought and research.[4] Expanding upon Foucault's archaeological approach, media are to be understood as the technological means to create cultural monuments and hence define on an operational level the law of what can be said. Hereby the media archaeological approach explicitly refers to Foucault's concept of the archive, which marks the goal of his methodology. Foucault himself was unaware of the media technological dimension of the formation and transformation of statements. This is not surprising as his inquiries end in the midst of the 19th century when the printed book and the painted image were still the dominant means of expression.

Since then new media technologies such as photography, film, typewriter, gramophone, radio, television, computer, Internet etc. have been developed. Consequently, the archaeological approach needs to address these medial changes in the way we express ourselves. In order to achieve this goal, we need to inquire media archaeologically, which is, according to Ernst, the task Foucault has left to us. Thus, the task is, as Wolfgang Ernst emphasizes, to study the forms of memory as operations of storing, processing, and transmitting information.[5] Hereby, the focus shifts from the factuality of actual statements to the virtuality of possible statements. The historical a priori of the archive in the sense of Foucault is thus no longer investigated in the institutional archives and libraries that preserved what has been said, but it is inquired as the material media technological means of information storage, transmission, and processing.

With the rapid shift from traditional analogue media to new digital media throughout the last 60 years, new possibilities as well as new challenges for the goal of preserving cultural heritage arose. As was already shortly pointed out, the idea of universal archives was reinforced with the advent of digital media. This idea lingers on until today and it resulted in startling projects that address the vision of universal archiving on different levels such as *MyLifeBits* that is run by Gordon Bell at Microsoft and the

Google Library Book Search Project. Yet, these attempts to create universal archives and libraries are, at least to some extent, undermined by the material constraints of digital media.

In 1998, Danny Hillis warned that we are at the verge of a "digital dark age", pointing towards the serious challenges we face with regard to our digital heritage.[6] Put simply, digital information does not last forever or, how Jeff Rothenberg put it, [d]igital information lasts forever, or five years, whichever comes first".[7] A large amount of digital information has already been lost. Day after day, hard drives and other storage media break down and the data stored on them often cannot or will not be recovered. The threat of data loss hence is a serious drawback of the glorious digital age.

In 2003, the UNESCO adopted the "Charter on the Preservation of the Digital Heritage" at its 32nd General Conference. Hereby, it is acknowledged that "resources of information and creative expression are increasingly produced, distributed, accessed and maintained in digital form".[8] As digital media become increasingly important, artefacts that are created by means of and for presentation with those technologies have to be recognized as "unique resources of human knowledge and expression"[9] and thereby become a crucial part of cultural heritage. It is a heritage, however, that is ephemeral in nature. The obsolescence of hard- and software contribute to this as well as the uncertainty regarding responsibilities, methods and legal implications of digital heritage. Digital preservation, thus, is acknowledged as a technological as well a societal and cultural problem that has to be addressed on two levels. First, digital artefacts have to be preserved for future generations and, second, accessibility of those artefacts has to be ensured.

3. Internet Archive's Utopia

Let us now turn from digital media in general to the Internet in particular. On the technological level, the Internet is an infrastructure or framework for computer-mediated interpersonal communication as well as computer-mediated publication of data, information, and knowledge. Conceptually, the Internet can be understood as a *docuverse*, that is, the sum of all documents and information posted on the net.[10] One of the basic characteristics of the Internet's content is its dynamic development. As every user is allowed to publish (nearly) anything he or she wants in the Internet and as everybody is able to change or remove this information freely, the web as a whole can be understood as a dynamic entity that, indeed, has a history of its own.

Early in the history of the WWW it became obvious that the life span of information accessible in the Internet is extremely short. Various numbers on how long information ordinarily lasts in the Internet can be found in scientific papers, magazine and newspaper articles, and online. According to Peter Lyman, the "average life span of a Web page is only 44 days, and 44

percent of the Web sites found in 1998 could not be found in 1999".[11] Ironically, Lyman had to admit that many of the sources he based his claim on had already disappeared from the web as well.

In 1996, a group of visionaries around Brewster Kahle founded the public non-profit organization Internet Archive. As the organization's website archive.org states, the main goal of Internet Archive today is to provide "Universal Access to All Human Knowledge".[12]. In this context, the Internet Archive attempts to preserve the entire Internet for future generations. This is probably what Internet Archive is still best known for today.

Archiving the entire Internet is by no means a trivial task and it is worth looking at the technologies and concepts that underlie the Internet Archive's attempt to preserve the Internet's past for future generations. What are the rules or practices that govern the passing on and transformation of online information those constitute the archive in the Foucaultian sense?[13] Thus, what can be found in the web archive of the Internet Archive? Using the so-called *Wayback Machine* on archive.org one can find snapshots of all web sites that are stored in the Internet Archive's web servers taken at different times in different intervals by typing in the desired URL and only that. These snapshots are created by a technology called web crawler, which is a software application that crawls through the net starting at certain entry points and following the hyperlinks on the websites to move from page to page and from site to site storing all the information retrieved on the servers along with some descriptive metadata. The first complete crawl through the Internet, according to Kahle, took about one year.[14] Today it still takes several weeks to compile a copy of the web.[15]

The Internet Archive takes a holistic approach to the web-archiving challenge. Instead of archiving only specific web sites that seem of historic importance, they strive to get it all. This is part of the Internet Archive's promise to preserve the Internet and not just the valuable web sites. This suggests that the archive is non-selective in what will be preserved and what not. They just get it all. However, is this really true?

Of course, it is not and Kahle admitted to this in a 1997 article about the Internet Archive published in Scientific American. He stated:

> The text, graphics, audio clips and other data collected from the Web will never be comprehensive, because the crawler software cannot gain access to many of the hundreds of thousands of sites. Publishers restrict access to data or store documents in a format inaccessible to simple crawler programs. Still, the archive gives a feel of what the Web looks like during a given period of time even though it does not constitute a full record.[16]

How reliable is this "feel of completeness" the Internet Archive gives us? Is it true that only those web pages are excluded that underlies the obstacles of restricted access, copyright law, and weird format? Again, the answer is no. So, what else cannot be found in the Internet Archive?

First, the Internet Archive preserves only what is called the surface web. Underneath what is visible at the surface of the web exists another web, consisting, on the one hand, of restricted web pages and, on the other hand, of databases that contain vast amounts of information that can be retrieved through web sites, but are not accessible as web sites. This would not be much of a problem if this so-called Deep Web would not contain most of the information in the Internet.[17] Furthermore, due to the rapid development of content management systems and weblogs, a great percentage of today's web content is generated dynamically from databases at the time of access. Thus, what a web crawler mirrors is only a small part of the possible web pages that could have been created from the database. The Internet Archive's approach to the web is document-oriented and everything that is not a document with a specific URL will not be included in the archive. As the Internet develops, more and more content is contained in databases and the Internet Archive becomes less and less reliable in giving us a feel of how the Internet was a at a certain time. In the age of Web 2.0 this becomes especially problematic, because most of the Web 2.0 services are database driven and only few of them follow the document logic imposed by the Internet Archive.

Second, web crawlers are software applications. Like all software applications, some work better than others do, but none of them is flawless. Programming always is a trade off between technological constraints and conceptual ideas. Thinking, for example, about the competition between search engines gives us a pretty good insight into the technological selectivity of web crawlers. In the late 1990ies and early 2000ies, a number of so-called *search engine wars* took place, and what the search engine providers were fighting about was how good their own crawlers searched the web.[18] One crucial factor for the usefulness of search engines is how many web pages they have indexed and how many pages hence possibly can be found. The web is smaller according to search engines that have fewer pages indexed. As every search engine provider uses web crawler applications like the Internet Archive we can safely conclude that their own crawlers underlie the same constraints. As a result, what future generations will remember of the Internet's past is technologically biased. To a certain extent this is due to conceptual decisions that are made when one opts to use web crawler applications to gather archival data.

As Mike Thelwall and Liwen Vaughan showed in 2004, a country imbalance exists in the Internet Archive. While 92 percent of the US websites were in the archive, there was only a probability of 70 percent for websites from China, Singapore, and Taiwan to be included in the Internet Archive.

Interestingly, Thelwall and Vaughan showed that this country imbalance was due to the biased link structures on the net.[19] That is, for regions of the web where web pages are only loosely interlinked it is less likely for these pages to be included in the Internet Archive.

4. Today's Memories of the Internet Archive's Past

After having briefly discussed the technological and conceptual means underlying the Internet Archive's attempt to preserve the entire Internet, I want to discuss the question what kind of memory we can gain by accessing the Web Archive. I will address this question by looking at the Internet Archive's memory of itself.

Being interested in the history of the Internet Archive, one might turn to their website to get some information on this topic. Yet, only little information on the history of the Internet Archive can be found there. The "About IA"-page just states that the organization was founded in 1996 and that it has broadened its scope in 1999 to not only preserve web pages but also texts, images, films, and software. The Internet Archive does not provide information on how it developed over time. That is, it does not tell its own history. However, why bother, if there is an archive containing the history? Converging with the *database logic* of the new media, the archive does not form a narrative of its own history in particular and of the Internet in general, it presents history as a list of items that can be looked at and compared.[20] Yet, turning to the Wayback Machine one finds only very little information on the history of Internet Archive as well and in this little information apparent contradictions.

Considering for example the question how the web archive grew in size over time, information was posted on the main page of the Archive between 1997 and 2001. During this time one could not only find information on the size of the archive but also information on which dates to which these numbers relate. Since then information regarding the current size of the web archive has vanished from the Internet Archive's main page and can only be retrieved from the FAQ section of the website. Furthermore, the information is not dated anymore. Reconstructing the growth of the web archive over time based upon these numbers reveals a surprising insight. Depending on when the Internet Archive's website was accessed contradicting information could be retrieved on how big the web archive was in March 2001. Figure 1 presents an estimate of the Internet Archive's growth drawing from (disagreeing) information by the Internet Archive itself. The graph was generated using information posted on the archive's website archive.org between 1997 and 2005.[21] Surprisingly, on March 31st, 2001 the website stated that the web archive had a size of approximately 42 terabytes at this time. Yet, on November 10th of the same year this information was changed, stating that the web archive had a size of 107 terabytes in March 2001 (hence

the diverging lines in 2001). This information retrieved from the Internet Archive's web archive is contrasted with a reliable claim by Arms et al. that the web archive had a size of 544 terabytes as of August 2005.[22] Yet, in August 2005 Internet Archive claimed on its website to have gathered already about one petabyte of web data, that is, about twice as much data as stated by Arms et al.

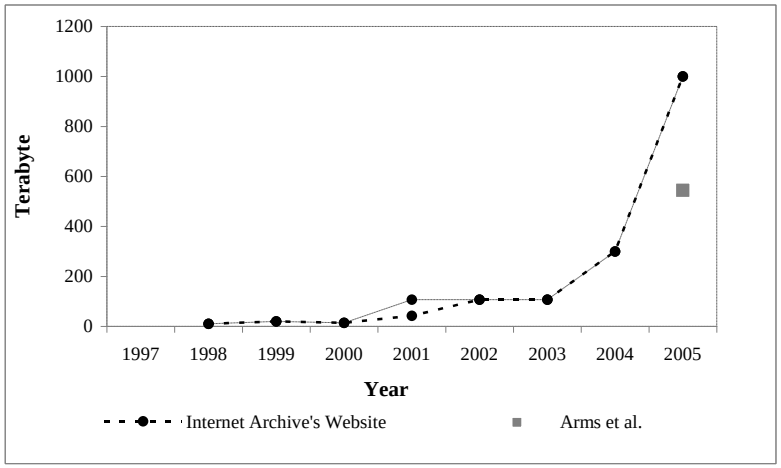

Figure 1: Growth of the Internet Archive's Web Archive

This sheds doubt on the reliability of information contained in the Internet Archive's web archive. Various explanations could explain these inconsistencies, but none of them can be preferred based upon the information in the web archive. Yet, there is a seemingly easy solution to the problem at hand. If we want to know how big the web archive was at a certain point in time, we cannot only rely on the information contained in the archive, but we could also gather information about the archive. By retrieving information on when crawls have been conducted and how much information has been gathered during this time, the growth of the web archive over time could easily be reconstructed. Yet, this leads to another problem. Until today, access to the web archive is only possible using the Wayback Machine that provides an interface, which allows users to retrieve the version history of a specific URL. A full text search of the archive or alternative ways of accessing the archives data are not provided to the general public. A more thorough access is promised for researchers from different fields upon request. Since December 2003, however, the Internet Archive is in the process of redesigning the researcher interface and, thus, cannot "process any new researcher requests".[23] Barriers to access pose an obvious hindrance to

our memory of the Internet's past. Without sufficient means to access the web archive our memory of the Internet is lost as well.

To conclude, starting with the general call for preserving digital artefacts for future generations I went on to discuss the Internet Archive's contributions to this goal. Indeed, the Internet Archive's collection is an invaluable resource for our future memory of the Internet's past. Taking a closer look at the technological and conceptual principles of the Internet Archive's web archiving project, challenges for preserving the entire Internet were discussed. After all it became clear that we might have to say farewell to the visions and promises of a universal archive.

Notes

[1] Anonymous, 'What Is Xanadu?', <http://www.xanadu.com.au>, viewed on 03.10.2009.
[2] Anonymous, 'Internet Archive', <archive.org>, viewed on 03.10.2009.
[3] M Foucault, *The Archaeology of Knowledge and the Discourse on Language*, Pantheon Books, New York, 1972.
[4] See for instance W Ernst, 'Medien@Rchäologie (Provokation Der Mediengeschichte)', in *Schnittstellte: Medien Und Kulturelle Kommunikation*, G Stanitzek and V K Wilhelm (eds), DuMont, Köln, 2001, pp. 250-67.
[5] W Ernst, 'Medien@Rchäologie'.
[6] See S Brand, 'Escaping the Digital Dark Age', *Library Journal*, Vol. 124, 2, 1999, pp. 46-48.
[7] J Rothenberg, 'Ensuring the Longevity of Digital Documents (Revised Version)', <http://www.clir.org/programs/otheractiv/ensuring.pdf>, 1999, viewed on 21.07. 2008.
[8] Unesco, 'Charter on the Preservation of the Digital Heritage', <http://portal.unesco.org/ci/en/ev.php-URL_ID=13367&URL_DO=DO_TOPIC&URL_SECTION=201.html>, 2003, viewed on 09.04. 2008.
[9] Ibid.
[10] T H Nelson, *Literary Machines*, Mindful Press, South Bend, 1982.
[11] P Lyman, 'Archiving the World Wide Web', *Building a National Strategy for Digital Preservation*, <http://www.clir.org/pubs/reports/pub106/pub106.pdf>, 2002, viewed on 10.10.2008, pp. 38-51, p. 38.
[12] Anonymous, 'Internet Archive'.
[13] Foucault, *Archaeology of Knowledge*, p. 127.
[14] B Kahle, 'Preserving the Internet', *Scientific American*, vol. 273, 3, 1997, pp. 72-73.

[15] W Y Arms et al, 'Building a Research Library for the History of the Web', in *Proceedings of the 6th Acm/Ieee-Cs Joint Conference on Digital Libraries*, 2006, pp. 95-102.
[16] Ibid.
[17] M K Bergman, 'The Deep Web: Surfacing Hidden Value', in *BrightPlanet.com*, <http://brightplanet.com/white-papers/119.html>, 2000, viewed on 10.11.2008.
[18] D Sullivan, 'Search Engine Sizes', <http://searchenginewatch.com/2156481#current>, viewed on 01.21.2009.
[19] M Thelwall and L Vaughan, 'A Fair History of the Web? Examining Country Balance in the Internet Archive', *Library & Information Science Research,* Vol. 26, 2004, pp. 162-76.
[20] L Manovich, *The Language of New Media,* MIT Press, Cambridge, 2001.
[21] Anonymous, 'Internet Archive'. The Wayback Machine was used to retrieve the main page of archive.org for the following dates to gather information the graph is based on: 10.11.1997, 10.13.1999, 04.08.2000, 03.31.2001, 11.10.2001, 05.27.2002. After 2002 information on the current size of the web archive appeared only in the FAQ section of the website. Data for 2003 to 2005 have been retrieved from archived versions of the FAQ page for the following dates: 09.08.2003, 06.09.2004, and 08.01.2005.
[22] Arms et al., 'Building a Research Library for the History of the Web', p. 96. Arms et al. are engaged in a project to create a research library using the data of Internet Archive's web archive. Hence, the authors had direct access to the data of the web archive and did not have to rely on information by Internet Archive.
[23] Anonymous, 'Internet Archive'.

Bibliography

Anonymous, 'Internet Archive', <archive.org>, viewed on 03.10.2009.

Anonymous, 'What is Xanadu?', <http://www.xanadu.com.au>, viewed on 03.10.2009.

Arms, W. Y., et al., 'Building a Research Library for the History of the Web', in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on DigitalLibraries*, 2006, 95-102.

Bergman, M. K., 'The Deep Web: Surfacing Hidden Value'. *BrightPlanet.com*, <http://brightplanet.com/white-papers/119.html>, 2000, viewed on 10.11.2008.

Brand, S., ‘Escaping the Digital Dark Age’. *Library Journal*, 124/2, 1999, 46-48.

Ernst, W., ‘Medien@rchäologie (Provokation der Mediengeschichte)’, in *Schnittstellte: Medien und Kulturelle Kommunikation*. G. Stanitzek and V. K. Wilhelm (eds), DuMont, Köln, 2001, pp. 250-67.

Foucault, M., The Archaeology of Knowledge and the Discourse on Language. Pantheon Books, New York, 1972.

Kahle, B., ‘Preserving the Internet'. *Scientific American*, vol. 273, 3, 1997, pp. 72-72.

Lyman, P., ‘Archiving the World Wide Web’. *Building a National Strategy for Digital Preservation*, <http://www.clir.org/pubs/reports/pub106/pub106.pdf>, 2002, viewed on 10.10.2008, pp. 38-51.

Manovich, L., The Language of New Media. MIT Press, Cambridge, 2001.

Nelson, T. H., Literary Machines. Mindful Press, South Bend, 1982.

Rothenberg, J., ‘Ensuring the Longevity of Digital Documents (revised version)’, <http://www.clir.org/programs/otheractiv/ensuring.pdf>, 1999, viewed on 21.07. 2008.

Sullivan, D., ‘Search Engine Sizes’, <http://searchenginewatch.com/2156481#current>, 2005, viewed on 01.21.2009.

Thelwall, M. and L. Vaughan, ‘A Fair History of the Web?: Examining Country Balance in the Internet Archive’. *Library & Information Science Research*, Vol. 26, 2004, pp. 162-76.

UNESCO, ‘Charter on the Preservation of the Digital Heritage’, <http://portal.unesco.org/ci/en/ev.php-URL_ID=13367&URL_DO=DO_TOPIC&URL_SECTION=201.html>, 2003, viewed on 09.04. 2008.

Marcus Burkhardt is a Ph.D. candidate at the International Graduate Centre for the Study of Culture at the University of Giessen, Germany. His main research interests are media theory and media philosophy, in particular the history and theory of digital media.