ASGER HARLUNG

ON ARCHIVED WEB

- NetLab 2016 -

Asger Harlung
On Archived Web
Version 1

Published by NetLab
Helsingforsgade 14, 8200 Aarhus N.

www.netlab.dk

Contents

1 Introduction and Purpose

Archived web means the preserved records of Internet-based human communication, primarily as represented in the form of websites. For practical reasons such records cannot be completely exact in covering all interlinked content, or capturing all ongoing changes. The ideal of preserving *everything* can never be realized completely: National libraries cannot capture and preserve all written content in all forms and versions; some will inevitably be missed. And with web archiving, data amounts, constant changes, and technical issues make preservation even more complex than with printed or other analogous content.

But archived web has become a necessary addition to national libraries and other forms of universal or research libraries; in preserving content and documenting changes and versions.

It is highly important for researchers who want to include studies of archived web to understand the possibilities and limitations of such web records.

This article will by no means offer a complete picture, but it will explain and highlight some basic principles of web archiving, and lay out examples and perspectives on research potentials of archived web. The article aims at explaining without being too technical for the average Internet user. The target group is researchers, students and anyone else interested in getting an overall feel for the perspectives of web archiving.

The article is by Asger Harlung. It was written in 2016, and uses contemporary data for explanations and examples.

2 Overview

In order to get an idea about what archived web is, and what the benefits (and some of the challenges) about archived web are, five major issues are central.

The article will present short introductions to each of them, but a general overview may be the first and final “sixth” issue:

1) What archived web means, and how and why web archives are created (see “Background and Explanation”).

2) The “big data” character of the web, and thus also of archived web (see “Abundance”).

3) That data cannot be complete (in web archives or anywhere else), and that archived web data differ from other types of archived media (see “Completeness and Authenticity”).

4) Searching in web archives may look like, but is significantly different from, searching the live web with online search engines (see “Searching a Web Archive”).

5) Types of use and research possibilities with archived web (see “Use and Perspectives”).

3 Background and Explanation

Archived web is the term for the data storage created by systematically storing content from the web. In daily language use, the terms “Internet” or “web” are often used as synonyms, but strictly speaking they are not. “Internet” means the sum total of technologies which connect computers in a worldwide network, while “web” means content created for and (mostly) by humans, shared and accessed via a network such as the Internet. In this article the word “web” is used in the sense of World Wide Web (WWW) and its domains.

Initiatives for archiving web content as part of the cultural heritage is a result of the same philosophy as that of State Libraries, National Archives, and other initiatives for creating universal libraries: The wish to preserve and gather as much information as possible, with the (impossible) ideal of gathering literally everything.

It is hardly necessary to explain the justifications and advances of universal libraries, where the history of ideas can be traced, near complete collections of thoughts on specific topics derived, lost ideas rediscovered, unchanged originals studied against revised versions, etc.

But since the emergence of the World Wide Web for the public in 1991, the World Wide Web has rapidly become the world’s single largest platform for creating and sharing content – in writing, sound, pictures, video, and software.

The urgent need for starting to archive web content soon became obvious. One of the most prominent initiatives is The Internet Archive founded in 1996. In Denmark, a national archive, Netarkivet, was founded in 2005.

4 Abundance

It would probably be fair to describe the amount of data moving and changing on the web as “staggering”.

For example:

- More than one hour of video is uploaded on Youtube.com every second. If someone should wish to analyse the contents of uploads from 2.5 hours it would take a full year, without pause or sleep, to see it all. And that is just for one major service.

- More than 10 million photos are uploaded to Facebook every hour.

- Google processes over 24 petabyte (one petabyte corresponds to 1015 characters; as a comparison 1015 seconds amounts to 37.5 million years) of data per day, which again, roughly, is over a thousand times the data amount stored in the world’s largest library; The Library of Congress in Washington, DC.

Add to such examples the fact that changes occur constantly and as abundantly as content is uploaded. Websites are born and die; and pages, comments and content are revised, moved around, repeated or deleted etc. constantly by millions of people online.

The undertaking of preserving as much as possible of what can be found on the web – and how it has been developed, changed and revised – is indeed a difficult and ambitious task. And also an important one, if anyone should wish to trace thoughts, development and dissemination of ideas and beliefs, public reactions to major events, in the place where such things primarily occur in the 21st century; which is the World Wide Web.

Web archives store not only millions of web pages, but also a multitude of copies of the same pages, preserving minor and major changes…

…or at least, a sort of overview of tendencies, because, naturally; if a website changes fourteen times and was archived four times over a year, then all changes are not documented. But at least, a sort of overview of changes over the year in question was preserved.

5 Completeness and Authenticity

It should be stressed from the beginning that web archives take various approaches towards protecting and respecting peoples' privacy. Data is only collected if it has been made public, and because data protection laws in Denmark assert that even then older data may be considered "personal", access to the Danish web archive, Netarkivet (http://netarkivet.dk/), is restricted to researchers with relevant projects, who must respect the rules and laws of privacy when working with the data.

It is also important to understand that data is collected on a regular basis, but that it is impossible to collect and preserve the amount of changes that occur (see also the previous subchapter, "Abundance"). Data in Netarkivet is primarily collected by automatically attempting to copy everything on the Danish National domain (.dk), and additionally websites relating to or located in Denmark four times per year. The process of data collection on such a scale is time-consuming and challenging, wherefore temporal differences may result in inconsistencies. For example: An event occurring during the time of data collection may be yet unknown on some pages stored at the beginning of the process, then announced or occurring on other pages stored underway, and finally commented upon as a past event on pages stored towards the end of the process.

Furthermore; most websites use content represented on other websites ("hosts"). For example: Video represented on Youtube can be a crucial part of a news article. But the data collection process will not gather the video; rather it will gather the code that should include the video on the news article. In some cases the video may be stored on the same website where the article is found, and in such cases it may be stored with the article. But in many cases the archived copy of the article will not contain the video, and will thus be incomplete, and if the video has been deleted or changed it may be difficult to fully reconstruct the original content of the article at all.

Finally, the process of gathering data holds its own challenges, and errors may occur, resulting in data incompleteness.

A detailed explanation of challenges and limitations is offered in Janne Nielsen's book "Using Web Archives in research – an Introduction, V2". It is a NetLab publication, and free to download, see "References".

The important thing is to understand that a web archive is not, and cannot be, an exact representation of the web as it was at a previous time. But it remains a large scale preservation of data which is changed over time and would otherwise be lost. With a web archive, patterns (of use, communication, user behavior or other phenomena) can be established with the degree of certainty that comes from observing them as big data patterns.

It may also be relevant to remember that no form of archiving can ever be complete. No amount of data can hold all information about a phenomenon, without actually being the phenomenon itself.

6 Searching a Web Archive

Search options offered in web archives are usually URL search (where the user searches for stored versions of a website by using its direct original web address, or “internet link”), and free text search. These are the search options offered in Netarkivet.

Free text search may appear similar to searching the live web with an online search engine such as Google. However, the reality of archive searches is quite different:

Several (often many) versions of the same website are stored, so that many “hits” in a search will be more or less identical. If changes were made to the different versions stored, then it may be difficult to determine which version best coincides with the one the user had in mind (if any).

Nevertheless, with the amounts of web data gathered over years, with many websites in many copies and versions, the amount of results from a free text search may be overwhelming. It will usually be in the user’s best interest to restrict a search as much as possible with web domain names, time limits, etc.

And even so, the user may face another problem which is that results may be listed alphabetically or chronologically, but not – as users of the live web are otherwise used to – by relevance. Online search engines use complex criteria (algorithms) to determine relevance and list the most relevant hits first. Offering results that are immediately perceived as relevant is crucial to the success of an online search engine.

But criteria of relevance depend on such things as popularity and exact representations of search phrases, which does not apply in a web archive in the same manner as online.

Web archive users may be accustomed to search results that are immediately useful, but even after using the best possible keywords and restrictions when searching in web archives the user may face results that will demand long and manual sorting before specific needs can be fulfilled.

Searching by URL also holds challenges, especially if the URL has been changed. The website may be there in earlier or later version than listed after a URL search, but under a different URL.

The chapter on searching in web archives in Nielsen (2016) is recommended for further reading, see “References”.

7 Use and Perspectives

The most common reason for users visiting a web archive is to retrieve a website or a web page which is no longer accessible online, or has been changed significantly.

This is a useful feature which may be very helpful, and similar to visiting a library in order to access a copy of rare or out of print material.

Another way to use web archives is as “big data”. Counts of hits for specific subjects, measurements of activity in relation to events, tracing activity patterns for specific subjects are obvious perspectives for use, and for research. (For example: “Where do debates on a subject occur; where do they link to, do they move from certain locations to other locations, etc.”).

And such use can be fully or partially automated in ways that would not be possible with data stored in an analogous form.

A combination of big data analyses and closer studies of selected or sampled examples will often be a good strategy for research purposes, combining solid observations of general trends with closer in-depth studies.

See also the page Ideas for examples of research possibilities using archived web.

8 References

Brügger, Niels (2015): Humaniora, Digital Humaniora, Medievidenskab og Internetforskning — en række mellemværender (Humanities, Digital Humanities, Media Studies, Internet Studies: An Inaugural Lecture), inaugural lecture, august 2015.

Available as

Sound (mp3, Danish language), or as

Text (pdf, English language).

(Figures mentioned in the article and in Niels Brüggers inaugural lecture are taken from:
Mayer-Schönberger & K. Cukier (2013): Big Data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt Publishing Company, New York, 2013, pp. 8-9).

Nielsen, Janne (2016): Using Web Archives in research – an Introduction, V2. Netlab, 2016.

A detailed introduction to web archiving for researchers, this is a free book from NetLab which can be downloaded here.

The book includes detailed descriptions of data collection (harvesting), challenges concerning completeness and authenticity, and search features in web archives and how to use them – each offering more detailed understanding of the basic principles laid out in the article “On Archived Web”.

Other noteworthy resources are:

Netarkivet (With access restricted to researchers).

The Internet Archive (Which can be accessed by the public).