ISSN 2278-5612

WEB ARCHIVING: TECHNIQUES, CHALLENGES, AND SOLUTIONS

[1]Adoghe Anthony, [1]Kayode Onasoga ,[1]Dike U. Ike, [1]Olujimi Ajayi
[1]Department of Electrical and Information Engineering, Covenant University, Nigeria.

ABSTRACT

Web archiving is the process of collecting valuable content from the World Wide Web in a an archival format, to ensure the information can be managed independently and preserved for the general public, historians, researchers, and future generation. If the Web is not preserved, eventually valuable content will be lost forever. The Web is a very valuable source of information and several government and private institutions are involved in archiving parts of it for various purposes. This paper gives an overview of web archiving, describes the techniques used in web archiving, discusses some challenges encountered during web archiving and gives possible solutions to these challenges.

Indexing terms/Keywords

Web archiving, World Wide Web, Researchers, Future generation, Historians.

Academic Discipline And Sub-Disciplines

Information Technology

SUBJECT CLASSIFICATION

Web Archiving

TYPE (METHOD/APPROACH)

Literary Analysis

Council for Innovative Research

Peer Review Research Publishing System

Journal: International Journal of Management & Information Technology

Vol. 5, No. 3
editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN 2278-5612

1.0 INTRODUCTION

The Internet is presently the bedrock upon which information is obtained. The most useful feature of the Internet is the World Wide Web, which is a unique information resource that hosts hundreds of millions (approximately 700 million as at 2012) of websites, that connects individuals, cities, and the world in general, using advanced web technology. The Web is in a constant state of evolution and there is no guarantee that its present content will remain usable in the distant future. Change in web content is caused by different reasons ranging from the personal inclination of the site owners towards the edition of parts of the content to accidental changes that occur during conversion into different formats. Even domain names are liable to change and omission. In recognition of this problem, national and private organizations from around the world have invested heavily in developing and applying the required technical tools to support more holistic web archiving solutions. An ever-growing international web archiving community continues to actively develop new tools to improve existing techniques, and to stop the continuous loss of web content caused by the transitory nature of World Wide Web. This paper provides an overview of the various web archiving techniques in use today, and the main challenges involved, and possible solutions.

2.0 OVERVIEW OF WEB ARCHIVING

The Web is the largest document ever written, with more than 4 billion public pages and additional 550 billion connected documents on call in the “deep” Web [1]. The Web keeps increasing in size, adding several million Web pages daily, and presently consists of billions pages of publicly available content and information. Content on these pages is structured in different ways, comes in of many formats and includes text, videos, and images as well as links between pages and to content in other formats such as PDF or docx. By using search engines and navigating through linked content users can find just about anything they are looking for. However, web pages are constantly being updated, relocated, or removed and you may not always be able to go back to something you saw before. For example, a video or article you found on a site last month may not be available anymore or the site you referenced in a paper may no longer even exist. Brewster Kahle, founder of the Internet Archive, once said that “the average life of a Web page is 100 days” [2]. Web pages disappear on a daily basis as their owners (authors) revise them or servers are brought out of service, users only find out when they enter a URL and get an error message “404 Site Not Found” [3]. Content is lost at an alarming rate, risking not just our digital cultural memory, but also organizational accountability [4]. To help preserve web content, websites are captured and archived for long-term access through web archiving.

Some organizations use simple tools and processes to archive their own web content. National libraries, national archives and various groups and organizations are also involved in archiving culturally important Web content in detail. Commercial web archiving software and services are also available to organizations that need to archive their own web content for their own business, heritage, regulatory, or legal purposes. The largest web archiving organization crawling the Web is the Internet Archive, an organization which aims to maintain an archive of the entire World Wide Web. The Internet Archive, a non-profit organization which aims to build a digital library of Internet sites, has been archiving websites and web content since 1999 with the purpose of preserving virtually “everything” and currently crawls portions of the web every few months [5], it is home to the largest collection of web content, with more than 2 petabytes of compressed data and over 150 billion websites captures [6]. Internet Archive manages to collect all types of websites extensively in order to maintain the powerful mirrors of Internet in the world. The tool which Internet Archive used to fulfill this purpose is Way-back Machine. Internet Archive archives websites by the topics, such as World War II, the US election, and so on. To allow libraries, archives, or content creators to create collections and archive selective web content, the Internet Archive created Archive-It, a subscription-based web archiving service. Web pages captured by the Internet Archive or through Archive-It are full-text and are searchable through their respective services. In addition to its own search and browse tools, the Internet Archive uses the Way-back Machine interface to allow users to search for specific URLs that may have been archived.

The concept of web archiving began at an initial stage in Taiwan. In contrast, Internet Archive, an unprofitable organization established at San Francisco, has been devoted to collecting all kinds of digital materials for the potential applications or researches since 1996. All kinds of websites are the targets of Internet Archive. National Library of Australia also built PANDORA and joined the web archiving in 1996 [7]. The United States of America, the United Kingdom, and Japan executed similar projects consequently.

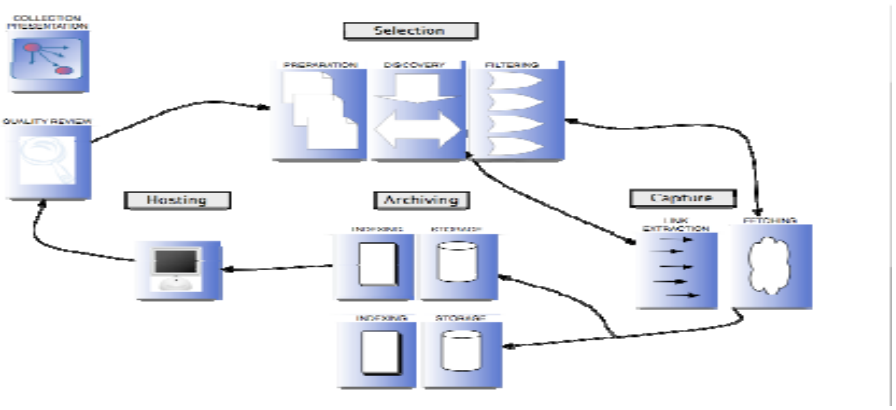

Fig.1: Web Archiving Process [19]

2.1 Reasons for Web Archiving

There are several reasons why we archive Web pages; a fundamental reason is to ensure both short and long term access to Web pages. More recent information is continually updated on websites which are added to the site and old pages are removed. However, the old content can still be a great value for capturing the site which can create history for the website and ensuring access to old content. Long term research is a usefully benefit from archiving a Web, as users have access to web content over time. Without the web archive, websites which change on daily basis may not provide long-term access to same content. The tracking down of references or resources whose original links have been removed is made possible via long-term access.

Also, as more and more of our records are digital, many things we do on a day to day basis are on the Web. Because of this it is sometimes just interesting to take a look back. Some web archiving projects have focused on this idea of creating time capsules for a glimpse into the past. When looking back at the older websites captured by the Internet Archive you can see how web pages in general have changed over time with innovations in technology. Record content is often subject to records laws, and as official records make the transition to digital form, agencies required to preserve and make this content accessible must still do so. Web archiving is one method that depository libraries and archives are using to fulfill this duty. Furthermore, many web archive collections are of government websites. In addition to capturing valuable information for the citizens, capturing government sites on a regular basis may help with government accountability or transparency as people will have to become accountable for their words and actions. Finally, archives provide an infrastructure with which information can be safeguarded for as long as necessary.

2.2 Web Archiving Techniques

Presently, there are several techniques used for the collection and preservation of websites, and these techniques vary according to the scale of the archiving operation. For large scale archiving operations, there are 3 major technical approaches [8]:

- Client side Archiving
- Transaction side Archiving
- Server side Archiving

Client-side archiving is the main Web acquisition method presently in use, this is because it is simple to use, cost-effective, scalable, and well adapted to a client–server environment. Web Crawlers such as HTTrack, Heritrix, and Wget act as clients, and make use of the HTTP protocol to collect content responses from the server. Crawlers start from seed pages (URL), parse them, extract links and retrieve the linked document. Each page has therefore to be “discovered” by link extraction from other pages. It is a powerful tool in the hands of the client and involves a web archiving crawler created and adapted from search engine technology. Figure 2, gives a representation of client-side archiving.

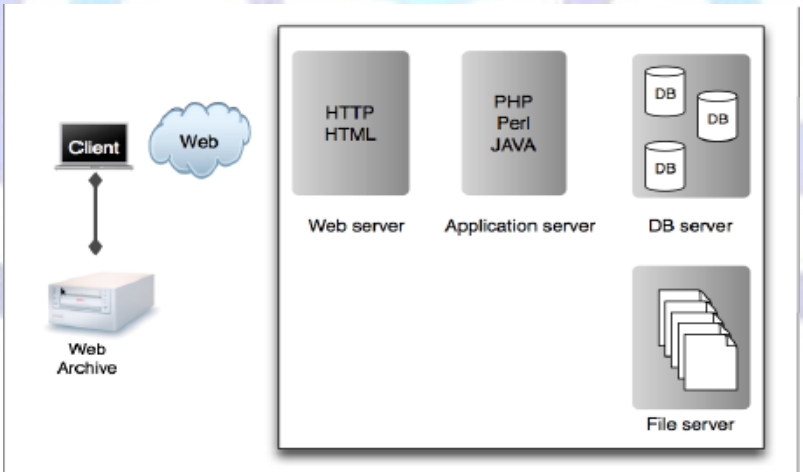

Fig. 2 Client side Archiving [20]

Transaction side archiving as the name entails, captures client-side transactions instead of hosted content and recording user access to website content based on client/server transactions. It involves the storage and archiving of Web content associated with all distinct HTTP response/request pairs. This is implemented in the Page-Vault system by using a filter into the Web server’s request and response flow. This type of Web archiving has the advantage of recording exactly what was seen and when. The main limitation of this method is the fact that it requires the use of code on the web-server that is hosting the content, and thus has to be implemented with the collaboration of the server’s owner. It is therefore used mainly for internal Web archiving by content owners. Figure 3, gives a representation of transaction-side archiving.

ISSN 2278-5612

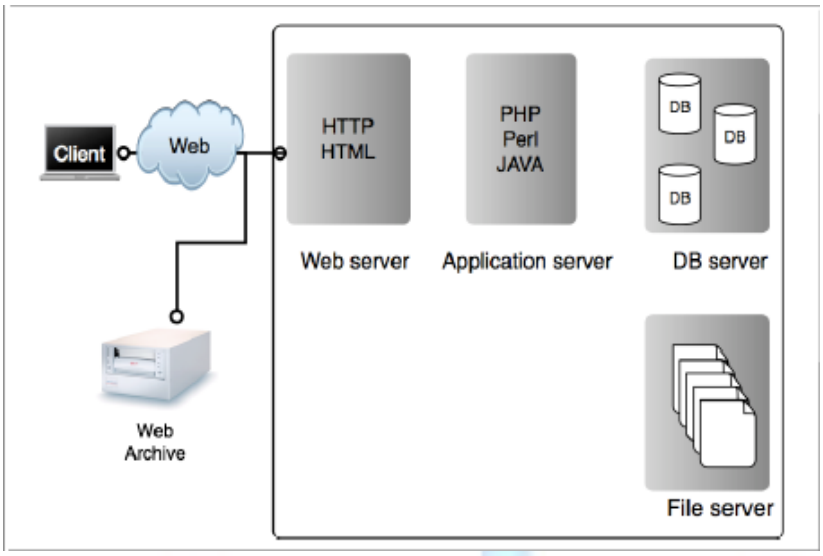

Fig. 3 Transaction side Archiving [20]

In server-side archiving, files are directly copied from the server without using HTTP protocol. Several kinds of information are obtained from servers generating a working version of the archived content. Its challenge is the back-up for its files. It can only be used with the collaboration of the site authors/owners. Although, it seems to be the most simple, it actually raises serious difficulty when generating a usable working version of the content, or when content is database driven. Nevertheless, it's a useful way to archive content otherwise missed by web crawlers. Figure 3, gives a representation of server-side archiving.

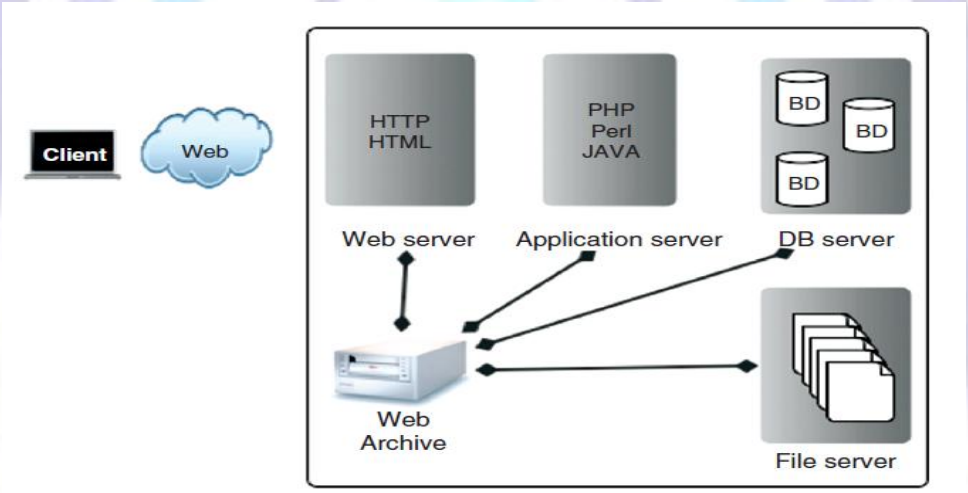

Fig. 4 Server Side Archiving [20]

Another approach uses RSS feeds to locate and pull externally hosted web content into a web archive, instead of using a web crawler. RSS feeds provide a trigger or alert that new content has been published, and can be used to ensure content is not missed in the intervals between periodic snapshot crawls [4].

Another important aspect of web archiving is the storage and organization of the web content. To this effect, three methods are employed for the organization and storage. These include:

- Local file system served archive
- Web served archive
- Non served archive

In Local File System served archive a local copy of the site file is created and navigated through in a way similar as on the World Wide Web. In web served archive, a web server is operated and its contents served in the environment of the user's browser. In non-served archive, documents are re-organized according to different logic naming, and addressing.

3.0 WEB ARCHIVING CHALLENGES AND SOLUTIONS

There are several reasons why the Web is difficult to collect and preserve. Some of these are technical, e.g. related to the size and nature of the Web itself [9], while others are related to legal, economic issues, and so on.

3.1 Legal Challenges

Legality is one of the most significant challenges to web archiving initiatives. Most web archiving organizations do not have the legal right to take copies of web content and provide access outside of the original site without the permission of the owner. Many websites display clear copyright information that disallows such activity. Also, there are potential problems with data protection, content liability and defamation. For example, in the UK, presently no single web archiving institution can harvest the entire UK domain without risking infringement in copyright. Though, the legal deposit legislation soon to be passed, will give UK legal deposit libraries the right to gather and provide access to all copies of websites published in the United Kingdom's web domain. The safest way of overcoming these challenges would be to select resources very carefully, thus excluding at source those resources that could have liability problems, and to develop effective rights management policies, combined with effective processes for the removal of certain types of material [9].

3.2 Economic Challenges

The economic problem is faced by all web archiving organizations. Since their mission is to provide web documents for many years, the return on investment (ROI) may be intangible, hence hard to quantify. Web archives require substantial initial investments to cater for technology, research and development, and it must be built to a considerably large scale if it is to save the entire web continuously. One way of solving this economic limitations, is for respective national governments to provide funding to private web archiving institutions, and/or to set up national web archiving institutions.

3.3 Limitation of Web Crawlers

Web crawlers encounter problems when harvesting contents from database driven web-pages, streamed multimedia files, script code, password protected content, Java script driven menus, and so on. These contents are commonly referred to as the "deep web". Also, crawler traps and limits on crawl size are significant set backs to the operation of web crawlers. Extensive research to this problem is necessary to address these limitations.

3.4 Quality Issues

Another issue affecting web archiving is the quality of web contents. In the World Wide Web there are different qualities of materials, some are of high quality, while others are of low quality. Each web page might "range from a few characters to a few thousand, containing truth, falsehood, wisdom, propaganda or sheer nonsense" [10]. A survey carried out in 2001 showed that while there was a general contentment with the World Wide Web as a valuable tool for research, many had significant concerns about reliability, accuracy and value of the information available [11]. A solution to this problem is for archiving institutions to ensure that only materials with reasonable quality are archived from the web, they should therefore have some criteria to use to measure quality of web contents.

3.5 Malware

Malware (malicious software), is software programmed by attackers to disrupt computer operation, collect sensitive information, or gain access to private computers [12]. It can appear in the form of scripts, active content, code, and other software [13]. Malwares include viruses, worms, ransom-ware, Trojan horses, key-loggers, root-kits, adware, dialers, spyware, rogue security software, malicious BHOs and other malicious programs; the majority of active malware threats are usually Trojans or worms rather than viruses [14]. Malware is a common feature in internet web pages, archiving these pages along with the malware, poses serious security risks. A solution to this problem is for archives to scan harvests, identify malware and delete them from their repositories. This should be done with very reliable anti-virus soft-wares to prevent false positives.

3.6 Technical Challenges

The World Wide Web at every instant continuously increases in size, the total amount of information on the 'surface Web' was once calculated to be between 25 and 50 terabytes in 2001 [15]. But this figures does not reflect the Web in its entirety, the size estimates of the Web only tended to count "static pages" that are freely accessible to search engines and Web users" [16]. A large number of web pages are not easily accessible; these are referred to as the "deep Web" and could be up to 400 to 500 times bigger than the surface Web [17]. Another technical problem has to do with the incoherence of collected web pages, this occurs if during the harvesting process of the web crawler (which could take several weeks), parts of the website have been updated and web content from the top of the seed URL no longer matches those at the lower levels, thus the resulting collection is not a coherent representation of the website. This problem can be solved by using tools that ensure temporal coherence in a website. Also, the protocols and standards utilized by some websites makes them very difficult to be archived. A solution to this problem is on the way, in the future with the 'Semantic Web', whereby information on the Web is given well defined formats, so that machines can understand it, and process it [18].

4.0 Conclusion

In this paper, we have reviewed the existing techniques used in web archiving and also analyzed the challenges encountered during web archiving. Web archiving technology has significantly advanced in the last decade. There are now set of tools and services which enable us to preserve important aspects of our online memory. Despite all this achievements, web archives still encounter significant challenges, more attention and funding are needed to develop tools that can increase the reliability of the web archiving process and that can support long-term preservation of web archives.

REFENECES

[1] Lyman, Peter, and Hal Varian. 2000. How Much Information? Available at: http://www.sims.berkeley.edu/research/projects/how-much-info/

[2] A. Arvidson, K. Persson, and J. Mannerheim. The kulturarw3 project - the royal Swedish web archiw3e - an example of 'complete' collection of web pages. In 66th IFLA Council and General Conference, 2000. Available at: www.ifla.org/IV/ifla66/papers/154- 157e.htm.

[3] Peter Lyman. 2012. Archiving the World Wide Web. Available at: http://www.clir.org/pubs/reports/pub106/web.html

[4] Maureen Pennock. 2013. Web-Archiving. DPC Technology Watch Report, 13-01 March 2013.

[5] Maria Halkidi, Benjamin Nguyen, Iraklis Varlamis, and Mihalis Vazirgianis. Thesus: Organising web document collections based on semantics and clustering. Technical Report, 2002.

[6] Ainsworth, S, AlSum, A, SalahEldeen,H, Weigle, M, and Nelson, M 2013 How much of the web is archived? arxiv.org/abs/1212.6177 (last accessed 9-08-2013).

[7] T. Haveliwala. Efficient computation of pagerank. Technical report, Stanford University, 1999.

[8] Masanés, J (Ed.) 2006, Web Archiving, 1st edition, Berlin: Springer.

[9] Michael Day. 2003. Preserving the fabric of our lives: a survey of Web preservation initiatives. Available at: bhttp://www.ukoln.ac.uk/metadata/presentations/ecdl2003-day/day-paper.pdf.

[10] Chakrabarti, S., Dom, B.E., Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A., Kleinberg, J., Gibson, D.: Hypersearching the Web. Scientific American, pp.44-52, June 1999.

[11] Herring, S.D.: Using the World Wide Web for research: are faculty satisfied? Journal of Academic Librarianship, pp. 213-219, 2001.

[12] Malware. Available at: http://www.techterms.com/definition/malware

[13] "An Undirected Attack Against Critical Infrastructure". Available at: http://ics-cert.us-cert.gov/pdf/undirected_attack0905.pdf.

[14] Microsoft. Microsoft active malware threats. Available at: http://ics-cert.us-cert.gov/pdf/undirected_attack0905.pdf.

[15] Lyman, P., Varian, H.R.: How much information? University of California at Berkeley, School of InformationManagementandSystems,Berkeley,Calif.Availableat:http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html

[16] Barilan, J., Data collection methods on the Web for info metric purposes: a review and analysis. Scientometrics, pp.7-32, 2001.

[17] Bergman, M.K.: The deep Web: surfacing hidden value. Journal of Electronic Publishing (August 2001). Available at: http://www.press.umich.edu/jep/07-01/bergman.html

[18] W3C Semantic Web Activity. Available at: http://www.w3.org/2001/sw/

[19] Marc Spaniol, Introduction to Web Archiving. May 28, 2009.

[20] Julien Masanes, Web Archiving: Issues and Methods, European Web Archive.