IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 3, March 2016

Web Archiving: Past Present and Future of Evolving Multimedia Legacy

Meenakshi Srivastava[1], Dr. S.K. Singh[2], Dr. S.Q. Abbas[3]

Assistant Professor, Amity Institute of Information Technology, Amity University, Lucknow, India[1]

Professor, Amity Institute of Information Technology, Amity University, Lucknow, India[2]

Professor, Ambalika Institute of Information technology, Computer Science Department, Lucknow, India[3]

Abstract: A continuous up gradation of tools and methods is going on for archiving the web in a better way. Although, most accepted practices and common model for web archiving is yet to evolve.

Keywords: Web archiving, Harvesting, Multimedia, Technological tool.

I. INTRODUCTION

Internet has evolved itself at an exponential pace and so has evolved its role in our life and society. It's not an exaggeration to mention that evolution of internet has made several technologies obsolete much before than they were expected to be. Evolution is a part and parcel of life. Any technological tool that does not evolve, improve or upgrade itself is static and would perish on its own. Realizing the fact that evolution is bound to happen, methodology of recording the historical data should be designed.

Since Internet is very dynamic and robust, hence elements of the multimedia content, data and information available change continuously. Interestingly if we look into the data or information available before internet became an integral part of our every part of life and made us mercilessly dependent upon it, we would realize that all the data and information has been stored systematically and is retrievable as and when required. Books irrespective of the number of editions they had can be retrieved from library, art work from galleries, objects of historical value from museum, important government documents are systematically filed in there Archive, Audio and video clips of events of past are archived. But we have failed miserably to keep a trail of evolving elements of the data information and multimedia content on internet.

Data available has evolved through up gradation of the basic data which would have been used very initially or the building block data , people working on the data have edited and upgraded it time and again as per there requirement, perception or the thought process and have unintentionally destroyed or overwritten the data which was available at starting. Creating a data and multimedia content archive which will help in preserving data and multimedia content step by step as it gets edited or upgraded will certainly be of help for people who require retrieving the data and multimedia content trail.

Considering the ephemeral nature of data and information available on internet, it becomes very important to design a mechanism which has an inherent nature of keeping a trail of entire evolution of multimedia content and data.

II. WORLD WIDE WEB (WWW)

Internet has grown enormously since 1990 and hence usage of WWW has also increased. An increase in usage of www has resulted in rise of information available online and hence the quantity of contents also pertaining to it. Increase in usage of internet exposed its weakness also and biggest weakness which was realized very soon was the volatility of the contents available on internet [1]. Web has huge data and information which is used by people of varied backgrounds who are professional and nonprofessional as well. Information published on internet did not have the estimated life span.

Furthermore multimedia content, data and information published on internet is digital and has no backup since the same is not archived, thereby is not available for future use. Since data, information and web pages change frequently without having a record, trail or archive of the same, so are not be available in future. Hence digital preservation of the multimedia data and systematic archiving of the same is very essential. It is very important not only to archive the multimedia data but also to archive it in such a way that same is easily retrievable when required. Also it needs to be seen that unnecessary archiving of the un required information is not there as it would delay entire retrieval speed latter on.

III. WEB ARCHIVING – EVOLUTION

Importance of web archiving was realized at a very nascent stage itself and continuous development of Web Archiving tools is being done. These tools are evolving steadily and consistently. However, most accepted practices and model with common consensus for web archiving are yet to evolve. There has also been rise in number of web archiving programs but institutions working on it are still in search of best technological tools which can perform desired web archiving. Parallel development of web technology, makes it difficult to proportionately develop web archiving tools. Besides lack of availability of a well-accepted tool of web archiving, it has also been observed that people are yet to realize the importance of web archiving. Furthermore considering the size and the extent of the excess of World Wide Web it is

imperative that several institutions have to work hand in glove for making web archiving a successful phenomenon.

IV. INTERNET ARCHIVE

Internet Archive [1] is among the initial steps which were taken for Web Archive in 1996. National Library of Australia and Sweden [Ref] and Internet Archive started collecting data and information available on web. Similar initiatives were taken by National Library of Norway, Denmark, Finland and Iceland [1] in 1997-1998. Although all the institutions were following a different agenda of their own but web archiving of the data was a common feature among all. Internet Archive has made an attempt to collect maximum of multimedia content, data and information of web world. Internet archive allows free access for upload and downloads to its collection of digitalized material which includes multimedia contents like web sites, software, applications, games, movies, music, books etc. [3]. As of May 2014 its collection topped 15pentabytes [3].Archive of NASA images was created after an agreement between NASA and Internet Archive and public were provided access to the images, Audio and Video collection of NASA in July 2008 [3]. Internet Archive has tied up with several similar institutions, libraries and museums like Brooklyn Museum, MusicBrainz, Metropolitan Museum, Libre Map Project to create a huge collection of digitized material for public access.

V. ARCHIVE-IT

In 2006 Internet Archive in partnership with thirteen other institutions launched Archive-It for web archiving. Archive-It became a platform to help its stake holder institutions to manage the digital data collection. Archive-It has been growing in terms of partners attached to it ever since and has 238 partners as on Jan 2013 [4]. Archive-It collection assigns multimedia content, data and information of web pages to a specific collection. Archive-It is a paid service and hence offers complete assistance to its partners. Archive-It helps its user to curate, scope and manage their data as per there requirement. Users have liberty of choosing how often and how far the data is crawled; specific contents can be excluded from being crawled which helps in filtering the multimedia content and data thereby ensures unnecessary data is not crawled, improves the speed of data retrieval at latter stage. Users of Archive-It also have an option of specifically choosing selected robot.txt on host to host basis. Archive-It also provides technical support to its users in the entire process and assists them with scoping issues. Partners of Archive-It have an advantage of getting the back up of their entire multimedia content and data which is not the case with General Archive. Archive-It also crawls Umbra and hence multimedia content and data from social media sites like Twitter, Flickr, Instagram, Facebook etc are also captured.

VI. INTERNATIONAL INTERNET PRESERVATION CONSORTIUM

International Internet Preservation Consortium (IIPC) was formed in 2003 at National Library of France with twelve institution being part of it at the onset. These institutions were in agreement of three years initially to fund and being part of the projects taken up by IIPC. Organizations from over forty five countries are part of IIPC now. National Universities, Museums, Libraries, Archives and Cultural Heritage Institutions all are part of IIPC now and IIPC encourages membership for enquires [5]. IIPC is an organization which has an endeavour of scaling up the practices, technological tools, standards and scope of web archiving which will assist in creation of International Achieves and will also ensure that nature of original internet content is stored indefinitely. IIPC is making an attempt to improve collaboration among international internet communities to provide broad access, global exchange, acquiring and preserving data and web information. IIPC is encouraging enhanced usage of web archives for research programs and cultural heritage. IIPC is working towards creating a rich collection of worldwide internet multimedia content, which can be securely archived and accessed by future generations. In attempt to accomplish its goal IIPC has established dedicated committees which are focused on developing various parameters pertaining to web archiving. Framework, Researchers Requirements, Access Tools, Metrics and Test Bed, Deep Web and Content Management are dedicated committees which are striving towards achieving the goal of IIPC. [10]

Dedicated group of Framework is responsible for developing standard and models of web archiving and also enhances better functioning and technical coordination among IIPC member libraries.[1]
Metrics and Test Bed group is responsible to develop and define metrics for collection and archiving of web content, it also evaluates the performance of technological tools involved in web arching.[1]
Access Tools group is responsible for initiative and tools which allow present and future access to web archive. [1]
Deep Web group are involved in developing processes and technological tools which archive web content that is not available to web harvesters. [1]
Content Management group is involved in developing tools to manage collection of web multimedia content, data and information. [1]
Researchers Requirement group works to estimate the requirement, expectations and the gap in the archiving of web multimedia content, data and information after discussion with experts.
WARC Analytical Tool, Heritrix Crawler, WARC archival standard are all products developed by IIPC.[5]

VII. PANDORA (PRESERVING AND ACCESSING NETWORKED DOCUMENTARY RESOURCES OF AUSTRALIA)

National Library of Australia in 1996 established PANDORA as national web archive for collection of Australia's Online Publication. PANDAS (PANDORA Digital Archiving System) was first available system to archive data and information [12]. PANDORA has now collaborated with state libraries of Australia and cultural collecting organisations Australian Institute of Aboriginal

and Torres Strait Islander Studies, the Australian War Memorial, and the National Film and Sound Archive.[8]. PANDORA has been able to achieve following:

- Archive of selected online Australian Publication acclaimed worldwide.
- Designing of exhaustive framework for collection and provision for long term access and retrieval.
- Policy development consistent research for preservation of online publication.

VIII. CHALLENGES

Collecting and archiving contents of a web site which is relatively smaller, simpler and does not deal with complex data is comparatively easier and straightforward. Contents can be downloaded from the server and stored which would be a very simple and will not require much technical expertise.[12]Users are amazed with the results that web can give but at same time they are ignorant about the importance of web archiving. Often web pages and web sites change, appear and disappear leaving no trails. The pace at which web has developed and changed continuously due to up gradation in technology, poses altogether new and different challenges. Web is huge and growing at a phenomenal pace with its dynamic characteristics. Web archiving multimedia content, data and information at larger scale for longer duration of time is a complex affair and requires expertise. Updating web archive on regular basis proportionately as the site develops or upgrades by understanding, identifying and maintaining the difference between various versions of the site is a tedious and complicated task.[12] Furthermore considering the dynamic nature of web it is very vital to segregate important data which needs to be archived and what needs to be left, as archiving unnecessary data will dampen the retrieval speed at latter stage.
Various external challenges which remain unaddressed yet, also persist like:

- Legal Aspects of Copyright, Data Protection, Defamation, implication of court order (On some incidence which was collected and archived) on archived data.
- Quality Management of multimedia content and data by ensuring all important multimedia contents and data have been captured and will be retrievable long term while unnecessary data has been excluded.
- Long Term sustainability to ensure multimedia contents and data will be available for download in future and will be compatible with the advancements of technology.
- Technological up gradation of web archive is in line with rapid evolution of Web technology.

IX. CONCLUSION AND FUTURE WORK

The study can be summarized and concluded with following observations and suggestions on scope of future work.

- Although all the efforts right now are focused at developing World Class Archive but International standard and parameters on Web Archiving are yet not defined.
- Requirement of Guideline to filter and archive important and necessary multimedia content and data only, so that irrelevant multimedia content and data is not stored and has no untoward effect on speed of data retrieval at latter stage.
- Quality Management to ensure that web archive harvest and capture what was actually planned.
- Continuous Research and Development of tools which are more reliable in terms of capturing all relevant information appropriately.
- Law pertaining to Copyright, Data Protection and Defamation need to be designed, considering the growing usage of web archive.

Web Archiving has come a long way but still lot needs to be done. To ensure that as we will provide books, publications, cultural items of heritage value of our time to our future generations, similarly we should ensure to develop tools capable of storing important relevant multimedia content, data and information for future generations in Web Archives. As well as A model needs to be developed which can easily navigate among various Web Archives and retrieves the desired multimedia content when required.

REFERENCES

1. Joao Miranda “Web Harvesting and Archiving” [On-Line] URL: http://web.ist.utl.pt/joaocarvalhomiranda/docs/other/web_harvesting_and_archiving.pdf
2. Web Archiving [On-Line] URL: https://en.wikipedia.org/wiki/Web_archiving
3. Internet Archive [On-Line] URL: https://en.wikipedia.org/wiki/Internet_Archive
4. Molly Bragg, Kristine Hanna “The Web Archiving Life Cycle Model”[On-Line] URL:https://archiveit.org/static/files/archiveit_life_cycle_model.pdf
5. IIPS[On-line] URL: http://netpreserve.org
6. Web archiving initiatives [On-line] URL: https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives
7. Pandora Archive [Online] URL: https://en.wikipedia.org/wiki/Pandora_Archive
8. PANDORA [Online] URL: http://pandora.nla.gov.au
9. Web-Archiving DPC Technology (Attached)
10. World Library and Information Congress: 71st IFLA General Conference and Council (2005). [On-Line] URL: http://www.ifla.org/IV/ifla71/papers/194e-Lupovici.pdf]
11. Catherine Lupovici. “The International Internet Preservation Consortium”.[On-Line] URL:http://iwaw.europarchive.org/05/lupovici.pdf
12. Maureen Pennock “Web-Archiving DPC Technology Watch Report 2013”. [On-Line] URL:http://www.dpconline.org/component/docman/doc_download/865-dpctw13-01pdf