INVITED PAPER

The History of Web Archiving

An explosive amount of data has been generated in the past two decades and shared by billions of people over the Internet, and this paper discusses the history and the current challenges of archiving massive and extremely diverse amounts of user-generated WWW data.

By Masashi Toyoda, *Member IEEE*, and Masaru Kitsuregawa, *Senior Member IEEE*

ABSTRACT | This paper describes the history and the current challenges of archiving massive and extremely diverse amounts of user-generated data in an international environment on the World Wide Web and the technologies required for interoperability between service providers and for preserving their contents in the future.

KEYWORDS | Digital preservation; web archiving

I. INTRODUCTION

The World Wide Web proposed by Tim Berners Lee in 1990 totally changed the way of publishing and broadcasting. An explosive amount of information has been generated in the past two decades and shared by billions of people over the Internet. Most web content is born digital. That is, it is originally created in digital form and its lifetime is rather short. The importance of preserving digital information was recognized in the late 1990s and various institutes, including the Internet Archive and national libraries, started archiving web content. In 2003, the United Nations Educational, Scientific and Cultural Organization (UNESCO) regarded digital materials as a cultural heritage, and raised the need for action to preserve this digital heritage [1]. This section describes those web archiving activities.

II. INTERNET ARCHIVE

In 1996, Brewster Kahle founded the Internet Archive [2] as a 501(c)(3) nonprofit in the United States to provide permanent access to historical collections of digital materials, and began to archive web contents donated from Alexa Internet and others. The Archive expanded its collections to include scanned books, music, and videos in 1999. In 2001, it provided a search interface. This was the so-called “Wayback machine” to access historical versions of archived web pages by specified uniform resource locators (URLs). It also developed a full text search engine, the so-called “Recall,” in 2003, which was eventually stopped in 2004. Recall could search archived web pages by keywords, and show temporal histograms of keyword occurrences in the Archive. In 2007, it was officially recognized as a library by the State of California. The Archive includes 2.4 petabytes of web contents along with 1.7 petabytes of books, music, and video collections as of December 2010.

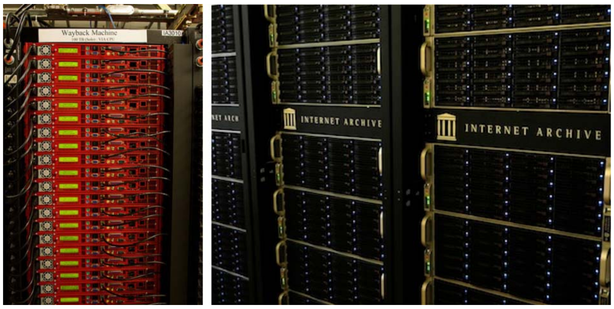

Fig. 1. ***Hardware of the Internet Archive (Petabox: http://www.archive.org/web/petabox.php).***

The Archive developed its own technologies for storing and collecting data. For storing petabytes of archived data safely, it designed a high-density, low-cost, and low-power storage system Petabox in 2004 (Fig. 1). For collecting web pages, the Archive developed open-source web crawler Heritrix with the Nordic national libraries from 2003, and it has been used in archiving projects by various organizations and national libraries.

Manuscript received January 16, 2012; accepted January 22, 2012. Date of publication April 11, 2012; date of current version May 10, 2012.

The authors are with Institute of Industrial Science, The University of Tokyo, Tokyo 113-0033, Japan (e-mail: toyoda@tkl.iis.u-tokyo.ac.jp; kitsure@tkl.iis.u-tokyo.ac.jp).

Digital Object Identifier: 10.1109/JPROC.2012.2189920

0018-9219/$31.00 ©2012 IEEE

III. INTERNET MEMORY FOUNDATION

The Internet Memory Foundation [3] was originally founded in 2004 as the European Archive, a nonprofit institution based in Amsterdam, The Netherlands and Paris, France. The foundation has been archiving dozens of terabytes of data per month in European region, and building access and navigation interfaces for archived data.

It has been involved in various projects funded by the European Commission to support growing and dynamic web archives. In the Living Web Archives project (LiWA) from 2008 to 2011, the foundation developed technologies including rich media capturing, temporal coherence analysis, spam assessment, and terminology evolution detection. These technologies were released as open source in 2010.

IV. NATIONAL WEB ARCHIVES

The most recent survey on web archiving initiatives in 2011 [4] identified 42 initiatives across the world including the above two initiatives. It shows that 80% of the archives focused on collecting information of their hosting country, region, of institution.

Most national and regional web archiving is performed by national libraries based on national acts for digital preservation. The earliest acts were in 2003 and 2004. In 2003, the National Library of New Zealand Act 2003 was established in New Zealand, and the Legal Deposit Libraries Act 2003 was established in the United Kingdom. In 2004, the Act to Establish the Library and Archives of Canada was established in Canada. These were followed by European and Asian countries.

In 2003, the International Internet Preserving Consortium (IIPC) was formally chartered with national libraries of France, Australia, Canada, Denmark, Finland, Iceland, Italy, Norway, Sweden, The British Library (U.K.), The Library of Congress (United States), and the Internet Archive. The IIPC has been developing common tools, techniques, and standards for building international archives. In 2008, it finalized the Web ARChive (WARC) file format based on the ARC file format used in the Internet Archive for archiving multiple web resources into one long file. The consortium also provides various toolkits for data acquisition, collection storage, and maintenance. The number of IIPC members significantly increased since 2003, and over 40 institutes participated as of 2011.

V. WEB ARCHIVES AT UNIVERSITIES

Several universities built their own web archives for research purposes. This section introduces the three largest archives: the Stanford WebBase Archive (Stanford, CA), the Socio-Sense system at the University of Tokyo (Tokyo, Japan), and the Web Infomall at Peking University (Beijing, China).

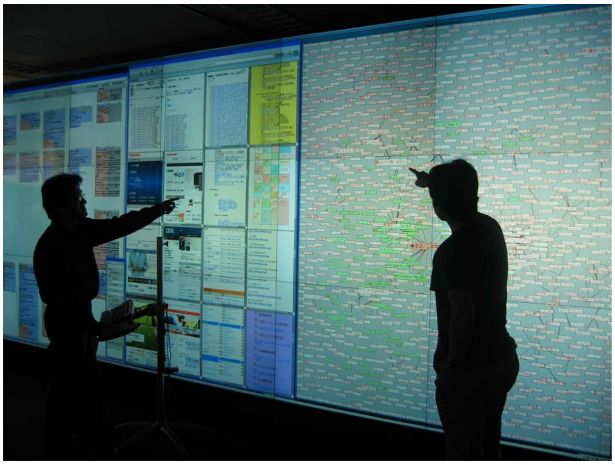

Fig. 2. ***The Socio-Sense System at the University of Tokyo.***

The Stanford WebBase Archive [5] was built as part of the Stanford Digital Libraries Project. The archive started its crawl from 2001, and has collected over seven billion web pages as of 2011. Along with general web crawls, it includes topic-focused snapshots of websites. For example, after the Katrina hurricane disaster, 350 sites were crawled every day for several weeks. The WebBase is used for developing technologies for crawling, designing web repository, and various search applications. Some earliest studies on parallel crawling, effective page refresh policies, and architecture for web repository are done on the WebBase. The archived data are opened to be explored by historians, sociologists, and public policy professionals.

The University of Tokyo is building the Socio-Sense system [6] for analyzing the societal behavior based on exhaustive web information, regarding the Web as a projection of the real world. The system consists of a huge-scale Japanese web archive, and various analytics engines with a large-scale display wall (Fig. 2). It has been crawling Japanese web pages incrementally for 11 years, and has content nearing 20 billion pages as of 2011. The Socio-Sense provides structural and temporal web analysis methods including web community mapping and extraction of their temporal evolution. The system has been used by sociological, linguistic, and marketing research experts.

The Peking University started to collect Chinese web pages in 2001, and has built the Chinese web archive, the so-called Web Infomall [7]. It has accumulated over three billion web pages as of 2011, and has been used for developing search engine technologies.

VI. CONCLUSION

The Internet includes various media formats such as images, movies, and interactive pages written in HTML5 and script languages. In addition, massive amounts of user-generated content tends to be isolated and concentrated into specific service providers, i.e., photo/video sharing services and social network services. Because of this situation, it has become more difficult to collect and preserve topical information. More sophisticated harvesting technologies for dynamic and multimedia content are required, and to establish international acts for promoting interoperability between service providers and for preserving their contents in the future. ■

REFERENCES

[1] UNESCO, "Charter on the preservation of digital heritage," 32nd Session General Conf. UNESCO, 2003. [Online]. Available: http://unesdoc.unesco.org/images/0013/001331/133171e.pdf

[2] Internet Archive. [Online]. Available: http://www.archive.org/

[3] Internet Memory Foundation. [Online]. Available: http://internetmemory.org/en/

[4] D. Gomes, J. Miranda, and M. Costa, "A survey on web archiving initiatives," in *Proc. 15th Int. Conf. Theory Practice Digital Libraries, Rese. Adv. Technol. Digital Libraries*, 2011, pp. 408–420.

[5] J. Cho, H. Garcia-Molina, T. Haveliwala, W. Lam, A. Paepcke, S. Raghavan, and G. Wesley, "Stanford WebBase components and applications," *ACM Trans. Internet Technol.*, vol. 6, no. 2, 2006, DOI: 10.1145/1149121.1149124.

[6] M. Kitsuregawa, T. Tamura, M. Toyoda, and N. Kaji, "Socio-sense: A system for analysing the societal behavior from long term web archive," *Progress in WWW Research and Development*, vol. 4976. Berlin, Germany: Springer-Verlag, 2008, pp. 1–8, ser. Lecture Notes in Computer Science, DOI: 10.1007/978-3-540-78849-2_1.

[7] H. Yan, L. Huang, C. Chen, and Z. Xie, "A new data storage and service model of China Web InfoMall," in *Proc. 4th Int. Web Archiving Workshop*, 2004. [Online]. Available: http://iwaw.europarchive.org/04/Hongfei.pdf

ABOUT THE AUTHORS

Masashi Toyoda (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from Tokyo Institute of Technology, Tokyo, Japan, in 1994, 1996, and 1999, respectively.

He worked at the Institute of Industrial Science, The University of Tokyo, Tokyo, Japan, as a Specially Appointed Associate Professor from 2004 to 2006. He is now an Associate Professor of Institute of Industrial Science, The University of Tokyo.

Dr. Toyoda is a member of the Association for Computing Machinery (ACM), IEEE Computer Society, Information Processing Society of Japan (IPSJ), and Japan Society for Software Science and Technology (JSSST).

Masaru Kitsuregawa (Senior Member, IEEE) received the Ph.D. degree in information engineering from The University of Tokyo, Tokyo, Japan, in 1983.

He is a Professor and the Director of the Center for Information Fusion at the Institute of Industrial Science, and Executive Director for Earth Observation Data Integration and Fusion Research Initiative (EDITORIA), The University of Tokyo. His research interests include database engineering and advanced storage system.

Dr. Kitsuregawa serves as the Chair of the steering committee of the IEEE International Conference on Data Engineering (ICDE) and has been a trustee of the VLDB Endowment. He was the recipient of the ACM SIGMOD E. F. Codd Innovation Award. He is serving as a Science Advisor to the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.