Blackwell Reference Online

Bibliographic Details

The International Encyclopedia of Communication

Edited by: Wolfgang Donsbach
eISBN: 9781405131995
Print publication date: 2008
Update: 2008-06-05 Revision History ↓

Archiving of Internet Content

STEVE SCHNEIDER AND KIRSTEN FOOT

Subject	Communication and Media Studies » Communication Studies Media System » Communication Technology, Internet and New Media
Period	2000 – present 1000 – 1999 » 1900–1999
Key-Topics	archives and documents
DOI:	10.1111/b.9781405131995.2008.x

Communication scholars interested in new media are increasingly archiving content of the → Internet or "web content," and studying web archives to examine retrospectively content produced and distributed on the web, and the behavior of those producing, sharing, and using the world wide web. Although web archiving has been actively pursued since the mid-1990s, new media scholars have more recently begun to find web archives helpful as they seek to understand developments related to the web in a variety of ways. These may involve adapting traditional methods of social research such as content analysis (→ Content Analysis, Qualitative; Content Analysis, Quantitative), ethnography (→ Ethnography of Communication), focus groups, → Surveys, and experiments (→ Experimental Design) or developing methods such as network ethnography (→ Network Analysis), hyperlink analysis, and other approaches to structural or phenomenological analyses of the web (→ Online Research).

RESEARCH HISTORY

The impetus for web archiving, for both scholarly and historical purposes, dates from the 1980s, as institutions increasingly shifted their records from paper to electronic form. With the advent of the world wide web in the early 1990s, this concern grew significantly, as greater numbers of institutions and individuals began producing documents in digital form only, rendering traditional modes of archiving less reliable as instruments to preserve records of social phenomena. By 1995, an opportunity to address this concern emerged with the development of *web harvesting programs*. Web harvesters or crawlers are applications that traverse the web following links to pages, initially from a set of pre-defined seed URLs. Harvesters were initially developed by search engines such as Alta Vista to overcome the increasingly impossible task of indexing the web through human cataloging techniques. Web harvesting technologies

gave life to the notion of archiving web materials via the web itself (→ Digital Media, History of; Internet, Technology of).

Three *pioneering efforts* to archive the web appear to have developed nearly simultaneously in 1995–1996: the Pandora Project by the National Library of Australia, the Internet Archive in the United States, and the Royal Library of Sweden. The impetus behind web archiving activity was clear: the web was doubling in size every three to six months from 1993 to 1996 and it appeared to many observers that it had the potential to become a significant platform on which a wide variety of social, political, scientific, and cultural phenomena would play out. Some individuals at institutions such as libraries and archives whose mission included the preservation of cultural and historical artifacts and materials recognized the challenge that digital materials presented and began developing techniques and institutions to address these challenges.

THE PROCESS OF WEB ARCHIVING

The construction of a web archive and accompanying infrastructure to facilitate scholarly analysis includes *four distinct processes*. First, the archive creators identify web pages and/or websites to be collected and specify rules for archiving page requisites, such as images, and for following links on archived pages. Next, crawling software is employed to collect the desired web objects (including html pages, images, etc.), and the archive creators verify that the crawl yielded an appropriate quantity and quality of archived objects. Analysis of archived objects – for example, to determine changes from one archived impression to the next – may then be performed, and metadata associated with archived objects can be created. Finally, a system for displaying archived impressions is developed and an interface created to provide access to the archive. The interface may support searching or facilitating selection of archived objects on the basis of metadata. Primary analyses may be completed using web archives created by scholars to examine specific questions. Secondary analyses use web archives created by other scholars, libraries, or archive organizations.

Each of the four processes of web archive construction poses significant technical challenges for researchers. Identification of pages or sites to be collected in an archive should be done in a systematic, clearly specified, and replicable process. If the objects in an archive represent a sample of web content relevant to a particular topic, the specification of the identification process is especially important to demonstrate the validity of the archive in relation to the pool of relevant web materials that were accessible at the time the archive was created. Further, specifications or rules regarding page requisites such as images and how links should be treated by the crawling software, entail additional challenges and have a significant impact on the number, types, and quality of pages archived and made available for scholarly analysis.

The actual collection of desired objects from the web, and subsequent storage of these objects on disks for future review and analysis, are relatively straightforward tasks. The challenge increases when archivists and scholars wish to be able to move from an archived page to other archived pages in hyperlink paths that were available when the archiving took place. This requires replicating the hyperlinked structure of the archived pages as it existed at the time of collection. Verification, especially in a large collection, is usually done at a macro level, based on some expectation of the total size of any given crawl. It is not often possible to physically review each archived page to insure that an accurate representation is available in the archive.

Once objects have been archived, archivists need to address the task of *generating and storing metadata* at a level of analysis appropriate to the anticipated research. Crawl programs generally produce log files documenting their processes which provide invaluable metadata about the archived objects. These log files need to be associated with the archived objects and considered as part of the archive itself. Archived objects can also be interrogated by software programs to develop additional machine-generated metadata. Further, scholars or analysts can be given an opportunity to create user-generated metadata about individual pages or sets of pages.

Finally, in order to make a web archive useful to scholars and other analysts, there must be a *system to search* for and display archived objects. Ideally, the search system will allow scholars to query the archive on the basis of all available metadata and to return a list of pages or objects matching the specified metadata. The system to display archived pages should clearly identify the source of the archived object and its archival date. Navigation to other archived impressions of the same object can also be provided.

ETHICAL AND OTHER CHALLENGES

In addition to this range of technical challenges, the processes associated with archiving web content yield significant and important ethical challenges as well. First and foremost, web crawling programs potentially impose costs on the

servers of the sites being crawled and on the owners of the sites being crawled. These costs – which can include bandwidth charges, denigration of service quality, loss of control over copyrighted information, and loss of privacy – ought to be taken into consideration by archivists. While a cost–benefit analysis may indicate that the benefits outweigh the costs, it should be acknowledged that the costs are not likely to be borne by the archiving entity. Scholars creating web archives may be under some obligation to insure that the archive is collected using crawling protocols that are acceptable to the producers of the sites being archived. For example, some institutions may require that producers of archived sites give permission to be crawled; others may restrict the distribution of archived pages to protect copyright and/or privacy (→ Research Ethics; Research Ethics: Internet Research).

In addition to creating their own web archives, some communication scholars also find value in web archives that have been created by others. There are additional challenges in working with the latter type of archive. Specifically, scholars working with web archives that have been collected as part of a general library or archive will need to address issues of sampling and boundaries, representation and selectivity, and archival standards.

The collection policy of an extant web archive, to the extent that it is available to the scholar, will provide an excellent window into the issue of sampling and boundaries. The range of sites collected and the criteria used to identify and specify the objects in the archive, will reveal a particular *sampling strategy*. In addition, the specification of which links to follow – and which to exclude from the archive – may be part of the collection policy. As all archives have boundaries or edges – representing the web space where pages or content linked from archived pages were not archived, the relevance of these boundaries to the representativeness of the archive should be examined. Scholars wishing to make generalizations based on sampled content should address these issues explicitly in their research.

Issues of representation and selectivity of archives are part of a somewhat broader set of issues associated with web archiving. First, there is growing concern that substantial archives of web content may *exclude significant portions of the web* – specifically, websites produced outside of North America, Europe, and Asia. This may become an increasingly salient issue for scholars interested in global issues and trends. In a related vein, it may be that archives will tend to include pages and sites that are highly ranked on prominent search engines – in a sense, replicating and perhaps reifying web traffic patterns while marginalizing less prominent websites. Finally, web archives will likely adhere to different standards, certainly in the near term while standards are being developed. Standards of verification, metadata, watermarking, and preservation may make research using web archives somewhat challenging.

Large institutions such as national libraries have engaged in both broad-based and thematic web archiving. In broad-based crawling activities, a web crawling program will start its archiving activity from sets of URLs that have no particular content relationship to each other. The web crawling program collects objects to be included in the archive by following links from the seed objects. This *automated collection process* continues indefinitely until all paths from the seed objects are exhausted or time and resources available reach their limit. A thematic web collection is an archive of web objects identified and captured using a set of URLs believed to be relevant to a specific theme or topic. A set of carefully selected URLs is used as the "seeds" for collecting activity. These URLs, representing either sites or pages of interest, are crawled at an established periodicity and with clearly specified rules concerning the crawling of linked pages and objects. For example, a crawler might be instructed to start at a given set of base URLs, to crawl all pages and page requisites with URLs from within the domain of the base URLs and to repeat this crawling procedure once per week for six months. These activities often result in archives with terabytes of archived objects. Scholars, especially those interested in studying the development of websites over time, may be particularly interested in thematic collections.

SEE ALSO: → Content Analysis, Qualitative → Content Analysis, Quantitative → Copyright → Digital Media, History of → Ethnography of Communication → Experimental Design → Internet → Internet, Technology of → Network Analysis → Online Research → Research Ethics → Research Ethics: Internet Research → Research Methods → Survey

References and Suggested Readings

Kahle, B. (1997). Preserving the Internet. *Scientific American*, (267) (3), 82–83.

Kavcic-Colic, A. (2003). Archiving the web: Some legal aspects. *Library Review*, (52) (5), 203–208.

Masanes, J. (ed.) (2006). *Web archiving*. Berlin: Springer.

Schneider, S., & Foot, K. (2005). Web sphere analysis: An approach to studying online action. In C. Hine (Ed.), *Virtual methods: Issues in social science research on the Internet*. Oxford: Berg.

Thelwall, M., & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology*, (57) (13), 1771–1779.

Cite this article

Schneider, Steve and Kirsten Foot. "Archiving of Internet Content." *The International Encyclopedia of Communication*. Donsbach, Wolfgang (ed). Blackwell Publishing, 2008. Blackwell Reference Online. 17 March 2014 <http://www.blackwellreference.com/subscriber/tocnode.html?id=g9781405131995_chunk_g97814051319956_ss51-1>

Copyright

Blackwell Publishing and its licensors hold the copyright in all material held in Blackwell Reference Online. No material may be resold or published elsewhere without Blackwell Publishing's written consent, save as authorised by a licence with Blackwell Publishing or to the extent required by the applicable law.