American Communication Journal
Vol. 9, No. 2, Summer 2007

Raising the Dead: Recovery of Decayed Online Citations

Daniela V. Dimitrova and Michael Bugeja

Key Words: *Internet research, online citations, Google, Wayback Machine, online archives, library research*

Recent studies show that online footnotes decay over time. This study investigates how researchers can resurrect lapsed citations, comparing two retrieval methods—online archives and search engines. The Wayback Machine and Google were used to retrieve missing online citations from six traditional communication journals. Analysis shows that the Wayback Machine was more efficient, suggesting archives are a better method for citation retrieval than search engines. Implications of these findings for scholars in multiple disciplines are discussed.

Daniela V. Dimitrova, PhD, is an Assistant Professor in the Greenlee School of Journalism and Communication at Iowa State University. Michael Bugeja, PhD, directs the Greenlee School at Iowa State University. Correspondence: Greenlee School of Journalism and Communication, 117 Hamilton Hall, Iowa State University, Ames, IA 50011
Email: DanielaD@iastate.edu

Introduction

While the Internet makes finding information easier and faster for researchers, it also often leads to problems with decaying online information. Recent studies have documented that the problem of "linkrot" is a serious issue not only for Web masters, but also for academicians who use Internet citations in their research. Scholars have found that between 10 and 40 percent of online citations tend to disappear from the Web (Rumsey, 2002; Markwell & Brooks, 2003; Sellitto, 2005; Tyler & McNeil, 2003; Sellitto, 2004). On average, about one-third of online citations vanish from the original Web location (e.g., Rumsey, 2002; Tyler & McNeil, 2003). This "half-life" phenomenon, or the time it takes for one-half of online citations to decay, has been observed across academic disciplines, including the area of mass communication (Bugeja & Dimitrova, 2005).

There are many reasons why online sources disappear. Some Web sites simply cease to exist because their Web site domain names expire. Other Web pages are removed from the Web by the Web site creator. Some Web sites are redesigned with new file structures. Others move to new online locations, bringing up a redirect message or going straight to the new URL, sometimes with different or updated content. Then there are server malfunctions and connection problems. Regardless of the reason for failure, most Web users have encountered the common error message "404: Page Not Found," a phrase that when entered in a search engine typically brings up a half-million or more entries on any given day.[1]

As more people heed the half-life issue, attempts have been made to alleviate the problem. While there is no easy solution, computer scientists and librarians are attempting to address the issue by working toward new persistent object identifier systems, one of which is the Digital Object Identifier (DOI) system. The DOI system will add sort of a bar code to each Web page so that it can be located using that code even when the page moves to a different URL (Lyons, 2005).

However, since there is no universally accepted technical solution as of yet, the authors of this study wanted to examine whether any reliable methods exist today to assist researchers in the recovery of vanished online citations. The goal of this study was to compare the reliability of two methods for citation retrieval—an online archive vs. an online search engine. The most popular Web archive—the *Wayback Machine*—and the most popular search engine—*Google*—were used for the purpose of retrieving missing online citations from six leading mass communication journals. We explored which method is more efficient and examined possible reasons why.

Finding Information in the Age of the Internet

Archiving

The Oxford dictionary (2006) online states that the noun "archive" means "a collection of historical documents or records." Archives of varied purposes and types have been in existence since primitive times when cultural histories were recorded via etchings in caves. In the 3rd millennium BC, records were stored on clay tablets in the Babylonian temple of Nippur, one of the earliest archives (Encylopaedia Britannica, 2003, p. 333). Clay tablets can be read to this day

[1] The phrase "404: Page Not Found" scored 589,000 on Google and 867,000 on Yahoo on 24 March 2006.

because sun-baked clay is almost imperishable (Harvey, 1987). Hence were developed the three criteria of archives, which remained unchanged until the digital revolution of the late 20th century: *place, implement,* and *material*. In the case of Nippur these included a temple, a quill to etch wet clay, and sun-dried tablets. In occidental culture the place that housed archives soon became known as the library, shifting from the Greek temples of the 4th century BC to the great ancient repositories of Alexandria and Pergamum (Encylopaedia Britannica, 2003, p. 333). Clay writing implements also were adapted to record content on papyrus, vellum, parchment, and paper.

The driving force behind such innovations was convenience, which historically has been of greater priority than permanence with respect to archival retrieval. Convenience has been associated with portability and storage while permanence has been associated with durability of material. Clay was more convenient to carry and contained more information than durable stone. In the same manner scrolls contained more data and were more portable than tablets; hand-written books were more portable and lengthier than scrolls; and printed books were more portable and lengthier still (Rychkov, 2003). However, what remained unchanged in archives from the mid 15th to the late 20th centuries were the components of place (library), implement (inked printing press), and material (paper). Moreover, the library owned the products of the printing press. All these factors have changed with the advent of the digital library, which exists in cyberspace and houses records owned by others that were created on software licensed by vendors and stored as files on servers.

Convenience still reigns supreme. The digital library can be "visited" at any time and from any place using portable technology to connect with cutting-edge storage systems vending information via license to users on demand. Librarians are finally taking note and "action to preserve online scholarly journals, saying they could vanish into oblivion should publishers go out of business or face other calamities," especially since libraries do not own and store content of journals licensed in electronic form (Foster, 2006). While librarians struggle to find a stable archiving system to store digital journals so that they do not vanish, another vital component inside those journals has been vanishing at an alarming rate: the footnote referencing online content. One factor has not changed since the time of Aristotle and Alexandria: Scholars still require reliable archives to retrieve durable content. However, since libraries increasingly no longer own the documents that they disseminate in digital form, researchers must rely on Internet-based repositories that exist in corporations or organizations rather than on campuses or in communities. This study investigates a small but vital aspect of that phenomenon.

The Wayback Machine

The most comprehensive Web archive—and possibly most reliable Web archiving tool today—is the *Wayback Machine* (http://www.archive.org/web/web.php).[2] As one prominent

[2] According to its Web site, "The Internet Archive *Wayback Machine* is a service that allows people to visit archived versions of Web sites. Visitors to the *Wayback Machine* can type in a URL, select a date range, and then begin surfing on an archived version of the Web. Imagine surfing circa 1999 and looking at all the Y2K hype, or revisiting an older version of your favorite Web site. The Internet Archive *Wayback Machine* can make all of this possible." Retrieved March 5, 2006, from http://www.archive.org/about/faqs.php#The_Wayback_Machine.

Web critic notes, it "is the closest thing to a permanent web archive" (Price, 2006). According to Dye (2005, 6), it is "a part of the Internet Archive, a nonprofit organization devoted to preserving data, texts, audio, Web sites, and other digital materials since the early days of the online revolution. Since 1996, the *Wayback Machine* has been sending out automated crawlers to all corners of the Internet and collecting digital, archived copies of everything they encounter." The *Wayback Machine* claims to include over 40 billion Web pages, which easily places it as the number one archive on the Web. Most Web archives outside of the *Wayback Machine* are smaller and for specific topics, such as elections, natural disasters, or scientific breakthroughs. The *Wayback Machine* and other popular archives can be found at http://www.archive.org. Special archives are designed to collect different types of media, such as images, video, and audio.

According to its Web site, the *Wayback Machine* began archiving Web sites in 1996. It uses Alexa Internet to crawl *only* publicly available Web sites. If someone doesn't want their Web site archived, they have the right to keep it from being archived. If someone's Web site isn't archived, they can request to have it archived for free. The actual archives are stored on petaboxes, which, as of right now, can hold one terabite of storage. It usually takes between six and twelve months to archive Web sites from the Alexa Internet crawler. Personal communication with one of the data archivists reveals that "Alexa crawls about 20-30TB per month. This rate has been increasing over the years. For example, we have only about 2TB of data from 1997, about 10TB from 1998, and the numbers grow from there" (Tofel, 2006). As more and more Web pages enter the World Wide Web, online archives need to store even larger amounts of data.

Search engines

Online search engines emerged in the early days of the Internet and have grown more sophisticated over time. Search tools, of course, preceded the World Wide Web. Most people can remember searching through the card-system of library catalogs. The paper of the cards and the oak drawers that they were stored in may not have been convenient, but they were durable. If the book or document was in the library, one could always find what one was looking for. As the card system went digital, with computer stations set up and accessible only in library facilities, a measure of convenience was added and a measure of durability was lost during computer crashes or updates. But the computerized search retrieval system that would migrate to the Web had been born. The engine that stored the most data became the most popular. Even though Yahoo, Altavista and HotBot are still in use, the most common search engine today is *Google* (Calishain & Dornfest, 2005). *Google* revolutionized the way online searches operate. As some observers state, "*Google* has changed the way people and computers alike approach the Web" (Calishain & Dornfest, 2005, p. xxvi).

Google

Google was established by two Stanford students, Sergey Brin and Larry Page, who are now its well-known founders (Hu, 2004). *Google* search engine began operation in its current form in 1998. It became popular very quickly. Now the term "google it" is accepted as synonymous with "search for it online." As *Google*'s online guide explains, its search has three distinct parts: a Web crawler that locates and fetches Web pages, an indexer for the retrieved online content, and a query processor that brings up the most relevant documents (GoogleGuide, 2006). One of the unique features of *Google*'s search algorithm is that Web pages that many users link to are ranked higher in the search results.

Without going into technical details, it is important to note that *Google* uses two different kinds of crawlers to fetch online content: deep crawl and fresh crawl. The first one probes deep within individual sites, making the indexing efficient. On the other hand, the fresh crawl is designed to index Web sites that change or update more often (i.e., online newspaper Web pages) in order to make the indexing more current. *Google* defaults to searching specified keywords, and that has become the methods that most people use and are familiar with now (Sullivan, 2005). It should be noted that *Google* offers a *cache* function, which brings a copy of the Web page when it was first indexed in the search engine. However, one needs more advanced computer knowledge to be able to use the cache function in Internet searches (i.e. "cache: www.yahoo.com").[3]

Research Question

Current research shows that, on average, about 30-40 percent of online citations vanish from the original URL (Bugeja & Dimitrova, 2005; Rumsey, 2002; Tyler & McNeil, 2003). Sometimes the decay rate is as high as 46% (Sellitto, 2004). This leaves scholars who want to find the original online citations with, arguably, two options: using an online search engine or a Web archive. If the half-life is mostly due to the removal of Web sites by the Web creator, search engines might not be the best way to find information since they only look for currently available sites; therefore, one might expect that online archives that save backup copies of Web pages may be better for finding original Web content. If the half-life occurs because online sources have been redirected or moved to new URLs, then search engines should be appropriate tools for locating these sources. To test these two different approaches, we arrive at our research question:

> ***Research Question 1:*** Which of the two methods—online archives or online search engines—performs better in reviving online citations used in the leading mass communication journals from 2000 to 2003?

Method

The goal of this study was to examine online citations given in articles in what are considered prestigious communication journals. We selected for analysis the following six journals: *Human Communication Research, Journal of Communication, Journalism & Mass Communication Quarterly, Internet Research, Journal of Broadcasting & Electronic Media,* and *New Media & Society.* These journals have high reputation as well as high impact factors according to the Institute for Scientific Information (ISI) Journal Citation Reports, now published by Thompson.[4] Details regarding each journal are provided in Appendix A. It is important to note that these are well established journals that are published in print form and then

[3] For example, using *cache:www.yahoo.com* in *Google* brings up the following message: "This is Google's cache of http://www.yahoo.com/ as retrieved on 8 Mar 2006 03:24:07 GMT. Google's cache is the snapshot that we took of the page as we crawled the web. The page may have changed since that time. Click here for the current page without highlighting. This cached page may reference images which are no longer available. Click here for the cached text only. To link to or bookmark this page, use the following url: http://www.google.com/search?hl=en&q=cache%3Awww.yahoo.com." Retrieved March 3, 2006, from http://www.google.com

[4] More details about ISI Journal Impact Factors can be found at the following URL: http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor.

made accessible via online databases. In general, the citations provided in these journals tend to point to refereed research articles although some of the online citation also point to research reports on Web sites such as the Pew Research Center for the People and the Press (http://people-press.org/), for example, or to federal government data such as the Census Bureau database (http://www.census.gov/).

Four publication years were chosen for analysis: 2000, 2001, 2002 and 2003. All research articles published during the four-year period in the six journals were downloaded from library databases and saved for analysis.[5] The unit of analysis was the individual URL for each online citation. All online citations that appeared in the research articles were examined to see whether the URL for the citation worked or not. If the URL worked, the variable was coded as 1 (Yes); conversely, if the URL did not work, it was coded as 0 (No).

This content analysis process yielded a total of 1,600 online citations from the six journals' research articles. Only 867 of the 1600 online citations (54.2%) were accessible in 2005 while 45.8% were inaccessible. The 733 "dead" URLs were of interest in our study. Two methods were used to test whether the dead URL could be found: first, using a search engine (*Google*) and then using an online archive (*Wayback Machine*) to retrieve the missing online citations.

The online search engine we chose was *Google*. As mentioned above, *Google* has become the most popular search engine today. Therefore, we assumed most people would first go to *Google* if they were trying to find a citation whose original URL had lapsed. Typically, when looking for an article, a person would go to the *Google* home page and type in the title and author name(s). The basic *Google* search has become a common search convention (Sullivan, 2005) and that was the approach we followed. *Google*'s Web search remains by far the most popular service that *Google* offers, accounting for about 80% of *Google*'s Web traffic, followed by *Google* image search with about 9% and Gmail with less than 6% of the traffic.[6] The traditional search engine's popularity is the main reason why we chose to use www.google.com as opposed to some of the more advanced search options.

Arguably, we could have chosen to use the *Google* cache option for citation retrieval.[7] As noted above, however, average users may not be familiar with the more technical options available via *Google*. Another possibility could have been to use *Google*'s service called *GoogleScholar*, which aims at providing only scholarly information, including journal articles, books, and cited in references link (Sullivan, 2004). However, this service is not considered as comprehensive as library databases (O'Leary, 2005). Another problem with *GoogleScholar* is the fact that it does not disclose the sources that it uses in the search indexing (Notess, 2005). Librarians do not recommend it for the above-mentioned reasons as well as for the fact that *GoogleScholar* is still in its beta version (not fully developed). Finally, the papers indexed in

[5] Only research articles were selected for analysis. Other types of publications such as editorial notes or book reviews were excluded.

[6] These data were collected in November 2005. See < http://blog.searchenginewatch.com/blog/051108-133720> for a detailed breakdown of *Google*'s services.

GoogleScholar determined by "unknown" procedure by *Google* typically focus on academic papers; however, the online citations we examined also pointed to research reports or company Web sites that are not classified as "academic." We did not want such citations to be excluded from the analysis.

The retrieval of the missing citations and their coding was conducted by a graduate student at a large Midwestern university. First, the coder was instructed to open up the *Google* home page at www.google.com. Next, the coder copied the article title and author name(s) as given in the citation source and hit the "Google Search" button. If the article was found via *Google*, the coder entered 1 (Yes). If the article could not be found, it was coded as 0 (No). In 5.6% of the 733 URLs, the URL did not reflect an article (i.e., there was no author name and title to be retrieved), which was coded as 9. It is important to note that when the search engine brought up results on several pages, only the first six pages of the search results were examined. Screen shots to illustrate the search process are provided in Appendix B.

The alternative method for retrieving the missing online citations was via an online archive. Again, we selected the most popular online archive, the *Wayback Machine*. The graduate student coder went to its home page, http://www.archive.org/web/web.php. The URLs of the vanished online citations were entered in the archive by copying the exact URL given in the online citation. Next, the coder clicked the "Take Me Back" button to search the online archive (See Appendix C for more details). If the URL was found via The *Wayback Machine*, the coder entered 1 (Yes). If the URL could not be found, it was coded as 0 (No). Several other coding categories were developed during the coding process: robots.txt query exclusion (coded as 2), blocked site error (coded as 3), and invalid request (coded as 4). These three error messages accounted for only 5.4% of the cases.

A second graduate student coder checked ten percent of the missing URLs in *Google* and *Wayback Machine*, following the procedures described above. Percentage agreement was established at 75.4% for *Google* and 100% for *Wayback Machine* retrieval. Average intercoder reliability for both variables of interest was 87.7%.

Results

Following the procedures outline above, a total of 1,600 online citations from the six mass communication journals were retrieved for analysis. Of particular interest here were the online citations that did not work when checked in 2005: 733 (45.8%) of 1,600 original citations. The 733 online citations that did not work—i.e., the citations that were inaccessible at the time this study was conducted—were first checked via *Google* and then checked via the *Wayback Machine*. The results from the two retrieval methods are compared below.

It is important to note that the coder also recorded whether the citations that were accessible in 2005 automatically redirected the Web visitor to a new URL. Slightly more than 21% of the working citations (184) did indeed take the visitor to a new Web location.

Google

The article title and author name(s) for each citation were entered into the *Google* Web search box and submitted to the search engine, as shown in Appendix B. The first six pages of the search results were examined. Only 201 (27.4%) of the sources were successfully retrieved following this procedure (See Table 1). Perhaps contrary to popular wisdom, more than 66% of the online sources cited in the six journals could not be found through *Google*. A small portion of the cited URLs (5.6%) did not have an author/title so they could not be entered into *Google*. Even smaller percentage of the searches (.8%) failed due to errors in the original dataset. Again, it is important to remember that only article title and author name(s) were entered in the search box and none of *Google*'s advanced search options was used, as explained above.

Table 1. Online citations retrieved through *Google*.

Citation Accessibility	Frequency	Percent
Citation found	201	27.4
Citation not found	485	66.2
Citation error	6	0.8
No author and title in citation	41	5.6
	733	100.0

The Wayback Machine

The URLs for the 733 inaccessible online citations were entered in the *Wayback Machine* search box and the "Take Me Back" button was clicked. More than half of the missing online citations in our database were revived using this procedure. The search in the *Wayback Machine* yielded a total of 392 (53.4%) Web sites of the original online citations that could be retrieved successfully in the year 2005 (See Table 2). We did not record how many times the Web site has been archived in the *Wayback Machine.*

Table 2. Online citations retrieved through the *Wayback Machine*.

Citation Accessibility	Frequency	Percent
Citation found	392	53.5
Citation not found	301	41.4
Robots.txt query error	34	4.6
Blocked site error	4	0.5
Invalid request	2	0.3
	733	100.0

More than 40%—301 URLs—had no matches in the *Wayback Machine.* Four URLs (.5%) brought up an error message within the *Wayback Machine*—"blocked site error"—while two (.3%) of the URLs that were entered were invalid requests. Interestingly, 34 of the 733 URLs (4.6%) brought up the following message: "robots.txt query exclusion." This message indicates that the crawler of online archive encountered a Web page using advanced scripting that could not be captured by the *Wayback Machine* crawler.

Comparison between **Google** *and the* **Wayback Machine**

Clearly, the *Wayback Machine* performed better in retrieving the decayed online citations from the six journals examined here: it resurrected almost twice as many of the online citations as *Google* did—53.5% and 27.4%, respectively. Of course, this is not a direct comparison since the online archive and the search engine have different primary functions and carry different types of content, as explained in the literature review.

We took a closer look to see whether the online citations found via the *Wayback Machine* were the same citations found via *Google.* Interestingly, there was an overlap of only 129 citations. In other words, 129 (64%) citations of the 201 citations retrieved through *Google* were also found in the *Wayback Machine* and only 72 citations were uniquely available through *Google*. In contrast, 263 (67%) of the 392 citations found in the *Wayback Machine* did not overlap and 129 (33%) overlapped with *Google*. The differences between *Google* and the *Wayback Machine* were statistically significant (Chi-square=12.75, p=.000, d.f.=1). Ultimately, if a researcher were to use both methods of citation retrieval, they would have been able to locate a total of 464 of the 733 original online sources or 63% of the above citations from the leading mass communication journals. If we go back to the original dataset and add the 464 revived citations, 269 (17%) of the 1600 original citations remain missing.

The breakdown of revived citations per year and per top-level domain (TLD) also reveals some interesting trends. Cross tabulations were run for article publication year and whether or not the citation was found in *Google* or the *Wayback Machine.* Interestingly, publication year did not emerge as a significant predictor of finding decayed online citations. Similar proportions of the revived citations came from each year. In contrast, TLD was a significant predictor for both *Google* and the *Wayback Machine* (Chi-square=57.37, p=.000, d.f.=4). The citations most likely to be found were from the *.org* domain: 81 (45.8%) of the citations found in *Google* and 123 (69.5%) of the citations found through the *Wayback Machine* had an *.org* Web address. In contrast, citations from the *.com* domain were least likely to be found via either retrieval method.

Figure 1. Percentage of recovered citations per year.

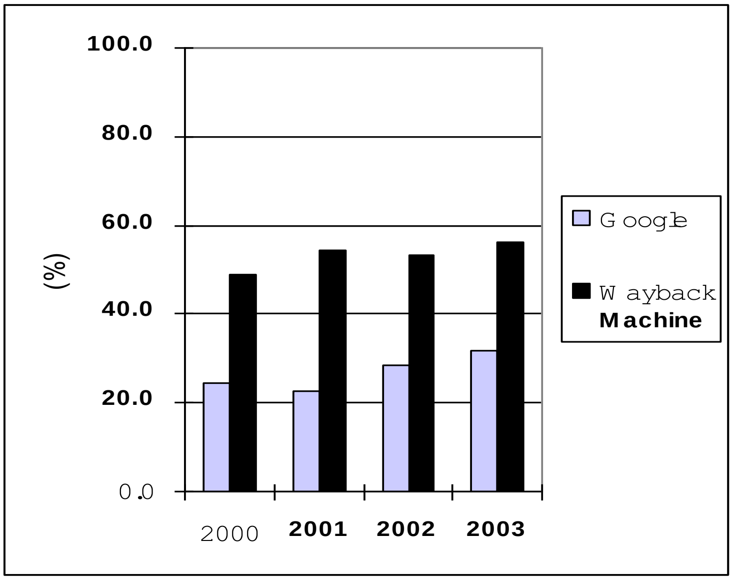
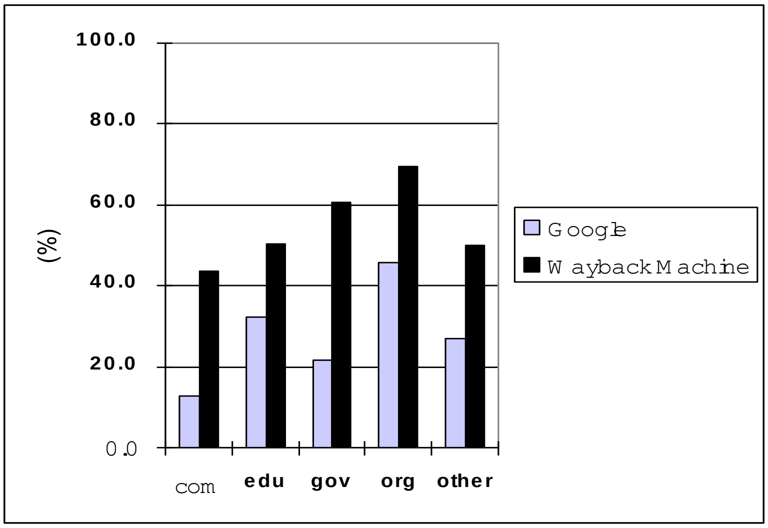


Figure 2. Percentage of recovered citations per top-level domain.



Discussion

The general findings of this content analysis suggest that authors trying to revive vanished online citations are better off using an online archive rather than a search engine at this point in time. This finding may be inherently related to the reasons why online information disappears: most likely structural reasons such as closing down of company or individual sites or moving Web sites to new addresses without providing a redirect link. However, the convenience of *Google*, combined with its immense storage capacity and popularity, are troublesome reminders that scholars may continue to use less reliable methods to retrieve lapsed citations. Moreover, unless graduate programs emphasize the importance of archives in their introductory research classes, the first impulse of newer researchers, trained in the digital rather than physical library, might be to check via search engine rather than archive.

Admittedly, however, there are limitations with both online searches and archives. For the most part online searches can only look for online content that is still available on the Web—in other words, content must endure and not disintegrate. The basic function of the online search crawler is to "crawl" in the metaphor of a spider snaring active content on the Web. If the content is dead, the chances of raising it again are limited, as this study shows. Conversely, the purpose of online archives is to collect Web sites and save them in a digital "museum" in case the URLs disappear or even update their online content. Therefore, it may not be as surprising that online archives perform better than search engines in retrieving missing online citations.

Online archives, however, have their own limitations (Bates, 2003). First, our study showed that the *Wayback Machine* can have internal search problems as evidenced by the error messages we came across: "Failed Connection" (which usually means that the server the archived material comes from is currently down); "Blocked Site Error" (which means that the site owner requested that their Web page stay out of archived material); and "Path Index Error" (which means that there is an internal error in the archived database for that particular page). *Wayback* also has limited access points since users have to type up the exact http address in order to find content in the archive (Bates, 2003).

Another limitation of the archive is the type of files it can capture. While the *Wayback Machine* has been successful at archiving html dynamic pages, it has had a hard time with JavaScript and Orphan pages.[8] There are frequent problems with image files. When images don't appear with a red "x" in their place on archived material, it means that the *Wayback Machine* simply didn't archive them.

Some of the good features of the *Wayback Machine* include the ease of use and free availability of archived Web material. Also, there are no organizational requirements to meet to access the archive. However, the *Wayback Machine* isn't a source of backing up a URL that goes down or is hacked, and doesn't guarantee that your page will be archived. The terms of use of the *Wayback Machine* state that archived material is not to be copied unless given authorization to do so. In fact, some companies concerned about copyright issues check the online archive on a regular basis to ensure their trademark logos are not captured (Kesmodel, 2005).

The World Wide Web is likely to grow further, which may create some capacity problems for online archives both in terms of storage of the archived Web content as well as in terms of crawling frequency to active online material. It is imperative to collect and preserve online content (text, audio, video, etc.) from a historical viewpoint since half-life is already a serious concern across academic disciplines (Dimitrova & Bugeja, 2006; Sellitto, 2005; Taylor & Hudson, 2000). The rate of decay of online content may even increase so future generations would need access to a "museum" of Web content to be able to retrieve data from the past. This feature of online archives has already been exploited in several lawsuits (Kesmodel, 2005) and even in online dating (Notess, 2004).

Conclusions and Implications

Before summarizing our conclusions, it should be noted that the International Internet Preservation Consortium (IIPC) is attempting to standardize online information retrieval and help reduce the half-life problem that directly impacts the integrity of digital archives, as this study has demonstrated. This group, which consists of national libraries from around the world, has a goal of developing standards for archiving, indexing, and serving content from the Internet. As this study also suggests, the challenge is great because the core components of stable archives—original sources owned by and housed in a real library—are in jeopardy in a digital age.

Conclusions and implications are as follows:

1. The most popular archive cannot locate about half of lapsed citations because its purpose is to resurrect dead Web pages rather than preserve information in a form of use for scholars. The implication here is the *Wayback Machine*, for all its attributes, is not a genuine archive in the library sense of that term. Scholars who perceive it to be might base citation on what is or is not available through *Wayback* rather than on what should or should not be cited in support of method or theory.
2. The most popular search engine, designed to find active content via key word rather than hyperlink, is almost twice as unreliable. *Google* is no archive, although its popularity and

[8] Orphan pages are pages that don't have any links to them. The Web crawler doesn't use search queries to find unlinked pages.

ease of use may suggest that it is. The implication here is that researchers unable to locate lapsed citations will turn to other less important sources whose primary distinction is online availability. The impact on research would be chilling in that a technical rather than scholarly attribute would dictate a substantial portion of content in journal articles.

3. The absence of a true digital archive for online citation threatens the very foundation of the scientific method: replication. If studies cannot be reproduced or be built upon by others because original sources cannot be found online, reliability of content in our leading journals will impact future studies on subjects of import to journalism and mass communication and, by extension, to society. The implication here further erodes the integrity and credibility of our disciplines, which have suffered from any number of ethical dilemmas stemming from increased Internet use in an age of convergence, including high-profile cases involving plagiarism and invention.

The combination of no true online archive, the continued popularity of search engines, and the erosion of the scientific method demand that research pedagogy in a digital age be re-examined on the basis of durability of retrieval over the long term rather than on the convenience of access over the short term. The implication of maintaining the status quo by utilizing digital libraries that license rather than own content may jeopardize research in mass communication. In sum, when citations become unstable and unreliable, by extension, our disciplines do, too.

References

Bates, M.E. (2003), "Archiving the web", *Online Medford*, Vol. 27 No. 6, p. 64.

Bugeja, M. and Dimitrova, D.V. (2005), "The half-life phenomenon: Eroding citations in journals", *The Serials Librarian*, Vol. 49 No. 3, pp. 115-23.

Calishain, T. and Dornfest, R. (2005), *Google hacks: Tips & tools for smarter searching,* O'Reilly & Associates, Sebastopol, CA.

The Compact Oxford English Dictionary. (2006), available at: http://www.askoxford.com/dictionaries/?view=uk (accessed 1 March 2006).

Dimitrova, D. V., & Bugeja, M. (2006). Consider the source: Predictors of online citation permanence in communication journals, *portal: Libraries and the Academy*, 6(3), 269-283.

Dye, J. (2005), "Web site sued", *EContent*, Vol 28 No. 10, pp. 6-7.

Foster, A.L. (2006), "Library leaders press colleges to archive online journals", *The Chronicle of Higher Education*, Vol. 52 No. 26, p. A33.

Google. (2006), "GoogleGuide", available at: http://www.googleguide.com/google_works.html (accessed 1 March 2006).

Harvey, E.H. (1987), *Book of Facts*, Pleasantville, New York, Readers Digest Association.

Hu, J. (2004), “Co-founders release Google ‘owner manual’”, available at: http://news.com.com/2100-1038-5202090.html (accessed 9 March 2006).

Kesmodel, D. (2005), “Not fade away -- Lawyers' delight: Old web material doesn't disappear; *Wayback Machine* and *Google* archive billions of pages, including deleted ones; Playboy protects 'sex court'”, *The Wall Street Journal Online*, available at: http://lists.essential.org/pipermail/ecommerce/2005q3/001909.html (accessed 9 March 2006).

Kesmodel, D. (2005), *Wall Street Journal (Eastern Edition)*, New York, N.Y.

Lyons, S. (2005), “Persistent identification of electronic documents and the future of footnotes”, *Law Library Journal*, Vol. 97 No. 4, pp. 681-94.

Markwell, J. and Brooks, D.W. (2003), “'Link rot' limits the usefulness of web-based educational materials in biochemistry and molecular biology”, *Biochemistry and Molecular Biology Education*, Vol. 31 No. 1, pp. 69-72.

The New Encyclopaedia Britannica (2003), Chicago, Encyclopaedia Britannica, (Vol. 7, p. 333).

Notess, G.R. (2005), “Dating the web: The confusion of chronology”, *Online Medford*, Vol. 28 No. 6, pp. 39-41.

Notess, G.R. (2005), “On the net: Scholarly web searching: *Google* Scholar and Scirus”, available at: http://www.infotoday.com/online/jul05/OnTheNet.shtml (accessed 9 March 2006).

O’Leary, M. (2005), “Google scholar: What's in it for you?”, *Information Today*, Vol. 22 No. 7, p. 35.

Price, G. (2006), “Google cached gone?”, E-mail newsletter to the members of NewsLib mailing list.

Rumsey, M. (2002), “Runaway train: Problems of permanence, accessibility, and stability in the use of web sources in law review citations”, *Law Library Journal*, Vol. 94 No. 1, pp. 27-39.

Rychkov, D.A. (2003), “Medieval manuscript production”, available at: http://library.rmwc.edu/hours/production.html (accessed 10 March 2006).

Sellitto, C. (2005), “The impact of impermanent Web-located citations: A study of 123 scholarly conference publications”, *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 7, pp. 695–703.

Sullivan, D. (2004), “Google Scholar offers access to academic information”, available at:

http://searchenginewatch.com/searchday/article.php/3437471 (accessed 9 March 2006).

Sullivan, D. (2005), "Google print is Google's ninth most popular service", available at: http://blog.searchenginewatch.com/blog/051108-133720 (accessed 9 March 2006).

Taylor, M.K., & Hudson, D. (2000), "'Linkrot' and the usefulness of web site bibliographies", *Reference & User Services Quarterly*, Vol. 39 No. 3, pp. 273-80.

Tofel, B. (2006), "Re: Requesting information", E-mail communication.

Tyler, D.C. and McNeil, B. (2003), "Librarians and link rot: A comparative analysis with some methodological considerations", *portal: Libraries and the Academy*, Vol. 3 No. 4, pp. 615-32.

The *Wayback Machine*. (2005), "Frequently Asked Questions", available at: http://www.archive.org/about/faqs.php#The_Wayback_Machine (accessed 1 March 2006).

Appendix A.

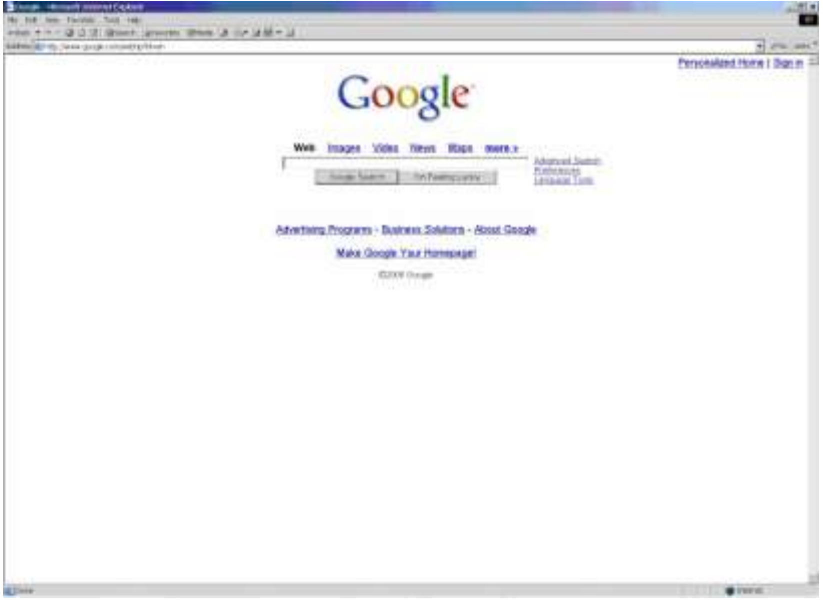
List of Journals and Journal Summary

Journal Title	Publisher	Year of First Edition	Frequency of Publication	Description
Human Communication Research	Blackwell Publishing	1974	Quarterly	ISI Journal Citation Reports® Ranking and Impact Factor: 2005: 8/42 (Communication), 1.080 2004: 1/40 (Communication), 1.526 *Human Communication Research* concentrates on presenting the best empirical work in the area of human communication. The journal has a broad social-science focus and as important applications to scholars in psychology, sociology, linguistics, and anthropology, as well as areas of communication studies. Topics include language and social interaction, nonverbal communication, interpersonal communication, health communication, intercultural communication, and developmental issues in communication. *Human Communication Research* is one of the official journals of the prestigious International Communication Association (ICA). (Source: http://www.oxfordjournals.org/humcom/about.html http://www.blackwellpublishing.com/journal.asp?ref=0360-3989)
Journal of Communication	Blackwell Publishing	1951	Quarterly	ISI Journal Citation Reports® Ranking and Impact Factor: 2005: 7/42 (Communication), 1.134 2003: 13/44 (Communication), 0.793 The *Journal of Communication* is a leading journal in the field of communication. Interdisciplinary in focus, *the Journal of Communication* concentrates on communication research, practice, policy, and theory. Since the Journal seeks to be a general forum for

				communication scholarship, it is especially interested in research whose significance crosses disciplinary boundaries. (Source: http://www.oxfordjournals.org/jnlcom/about.html http://www.blackwellpublishing.com/journal.asp?ref=0021-9916)
Journalism & Mass Communication Quarterly	Association for Education in Journalism and Mass Communication	1924	Quarterly	*No ISI Ranking *Journalism & Mass Communication Quarterly* focuses on research in journalism and mass communication. Articles report results from original investigation, presenting latest developments in theory and methodology of communication, international communication, journalism history, and social and legal problems. (Source: http://www.aejmc.org/pubs/#jmcq)
Internet Research	Emerald Group Publishing, Ltd.	1991	Five issues per year	ISI Journal Citation Reports® Ranking and Impact Factor: 2004: Listed, the total cites totaled 203 with an impact factor of 0.562. *Internet Research* is an interdisciplinary journal that publishes Internet-related research and aims to foster understanding of telecommunication networks in society. In addition to looking at the technological developments which facilitate their increasing use, this journal also examines the social, ethical, economic and political implications which result from public access to a wealth of information. (Source: http://www.emeraldinsight.com/info/journals/intr/intr.jsp)
Journal of Broadcasting & Electronic Media	Lawrence Erlbaum Associates, Inc.	1956	Quarterly	*No ISI Ranking The *Journal of Broadcasting & Electronic Media* is a scholarly journal published quarterly by the Broadcast Education Association (BEA). It is considered one of the leading publications in the mass communication field. *JoBEM* contains timely articles about new developments, trends and research on electronic media written by academicians, researchers and electronic media professionals. (Source: http://www.beaweb.org/jobem/info.html)
New Media & Society	Sage	1999	Six issues per year	ISI Journal Citation Reports® Ranking and Impact Factor: 2005: 14/42 (Communication), 0.855 *New Media & Society* is an international journal that provides an interdisciplinary forum for the examination of new media and social change. The journal publishes articles that are multidisciplinary in focus and come from both the social sciences and the humanities including areas such as media and cultural studies, sociology, geography, anthropology, and economics. (Source: http://www.sagepub.co.uk/journalsProdDesc.nav?prodId=Journal200834

Appendix B.
Sample query via *Google*

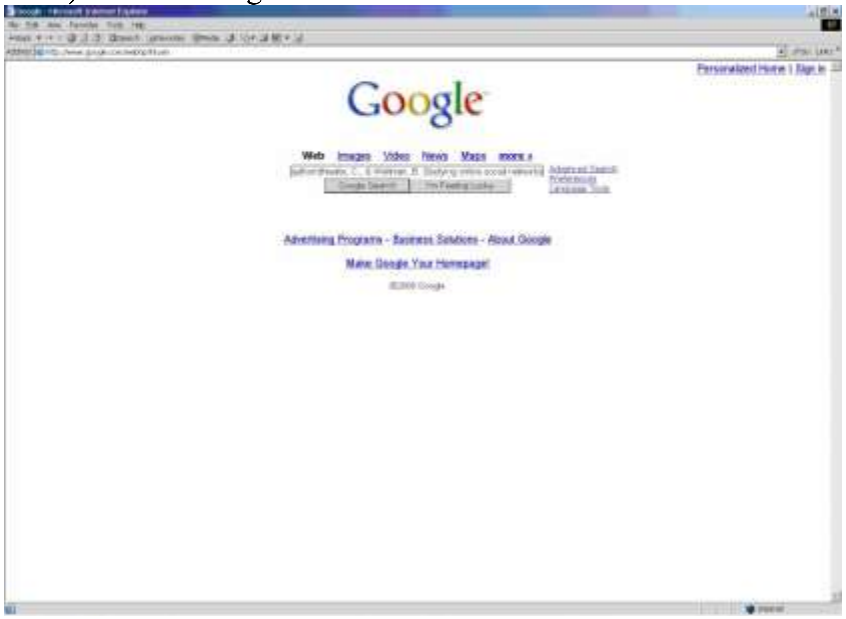
1) Access http://www.google.com.


Personalized Home | Sign in
Google
Web Images Video News Maps more »
Google Search
I'm Feeling Lucky
Advanced Search
Preferences
Language Tools
Advertising Programs - Business Solutions - About Google
Make Google Your Homepage!
©2006 Google

2) Type the name of author(s) and the title of the article to retrieve in the search box.

(In this sample query, we retrieved an article titled "Studying online social networks" by Garton, L., Haythornthwaite, C., & Wellman, B., originally located at http://jcmc.huji.ac.il/vol3/issue1/garton.html)

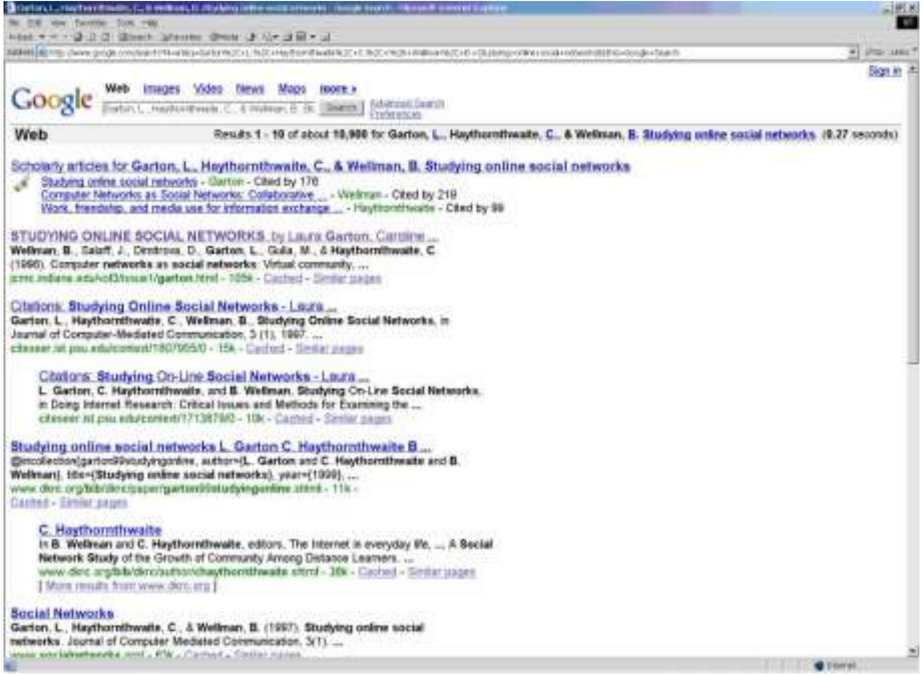
3) Click "Google Search."


aythornthwaite, C., & Wellman, B. Studying online social networks

Personalized Home | Sign in
Google
Web Images Video News Maps more »
Google Search
I'm Feeling Lucky
Advanced Search
Preferences
Language Tools
Advertising Programs - Business Solutions - About Google
Make Google Your Homepage!
©2006 Google

4)

5) The screen shot shows the result of the search.



Web Results 1 - 10 of about 10,900 for Garton, L., Haythornthwaite, C., & Wellman, B. Studying online social networks. (0.27 seconds)

Scholarly articles for Garton, L., Haythornthwaite, C., & Wellman, B. Studying online social networks
Studying online social networks - Garton - Cited by 176
Computer Networks as Social Networks: Collaborative ... - Wellman - Cited by 219
Work, friendship, and media use for information exchange ... - Haythornthwaite - Cited by 99

STUDYING ONLINE SOCIAL NETWORKS, by Laura Garton, Caroline ...
Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., & Haythornthwaite, C. (1996). Computer networks as social networks: Virtual community, ...
jcmc.indiana.edu/vol3/issue1/garton.html - 105k - Cached - Similar pages

Citations: Studying Online Social Networks - Laura ...
Garton, L., Haythornthwaite, C., Wellman, B., Studying Online Social Networks, in Journal of Computer-Mediated Communication, 3 (1), 1997. ...
citeseer.ist.psu.edu/context/1807955/0 - 15k - Cached - Similar pages

Citations: Studying On-Line Social Networks - Laura ...
L. Garton, C. Haythornthwaite, and B. Wellman. Studying On-Line Social Networks. in Doing Internet Research: Critical Issues and Methods for Examining the ...
citeseer.ist.psu.edu/context/1713879/0 - 10k - Cached - Similar pages

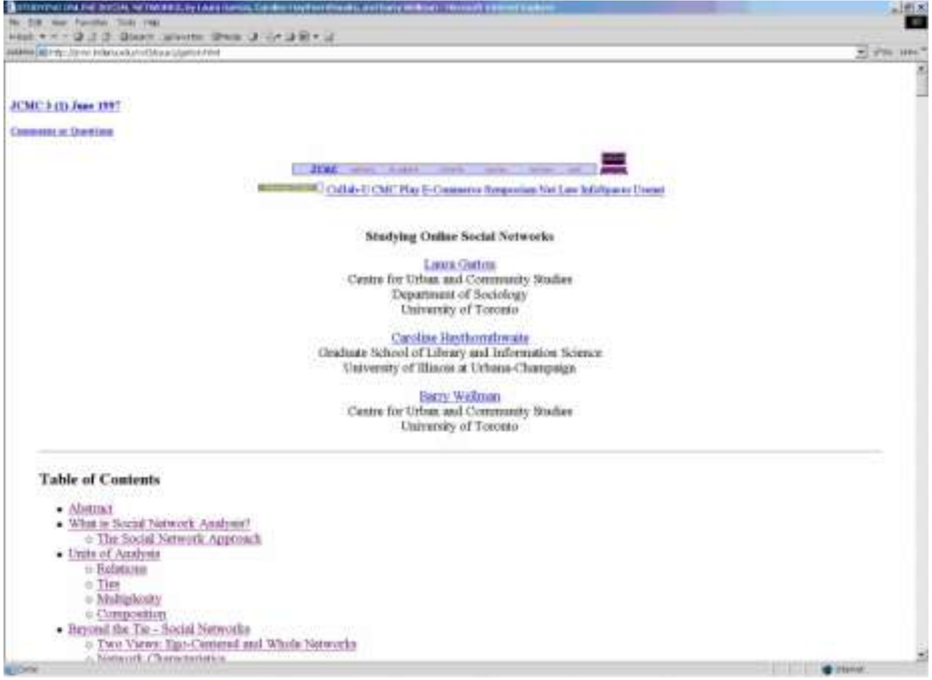
Studying online social networks L. Garton C. Haythornthwaite B ...
@incollection{garton99studyingonline, author={L. Garton and C. Haythornthwaite and B. Wellman}, title={Studying online social networks}, year={1999}, ...
www.dirc.org/bib/dirc/paper/garton99studyingonline.shtml - 11k - Cached - Similar pages

C. Haythornthwaite
In B. Wellman and C. Haythornthwaite, editors, The Internet in everyday life, ... A Social Network Study of the Growth of Community Among Distance Learners. ...
www.dirc.org/bib/dirc/author/chaythornthwaite.shtml - 38k - Cached - Similar pages
[More results from www.dirc.org]

Social Networks
Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. Journal of Computer Mediated Communication, 3(1). ...
www.socialnetworks.org/ - 87k - Cached - Similar pages

6) Find the matching article by clicking and checking the hyperlinks which look relevant.

7) Original article has been retrieved.


JCMC
Collab-U CMC Play E-Commerce Symposium Net Law InfoSpaces Usenet

JCMC 3 (1) June 1997

Comments or Questions

Studying Online Social Networks

Laura Garton
Centre for Urban and Community Studies
Department of Sociology
University of Toronto

Caroline Haythornthwaite
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

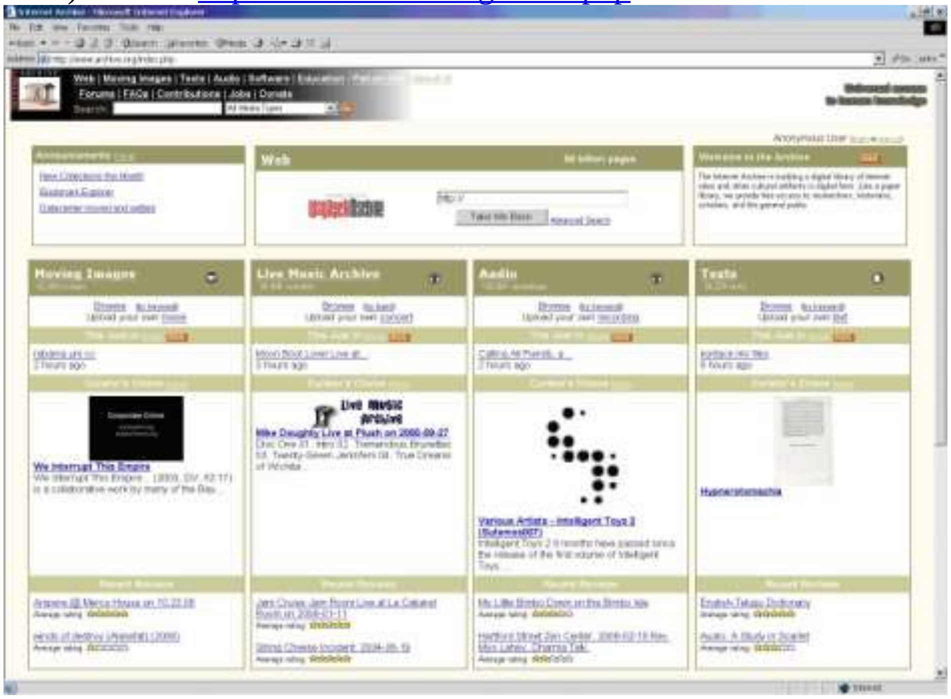
Barry Wellman
Centre for Urban and Community Studies
University of Toronto

Table of Contents

- Abstract
- What is Social Network Analysis?
  - The Social Network Approach
- Units of Analysis
  - Relations
  - Ties
  - Multiplexity
  - Composition
- Beyond the Tie - Social Networks
  - Two Views: Ego-Centered and Whole Networks
  - Network Characteristics

Appendix C.
Sample query via *Wayback Machine*

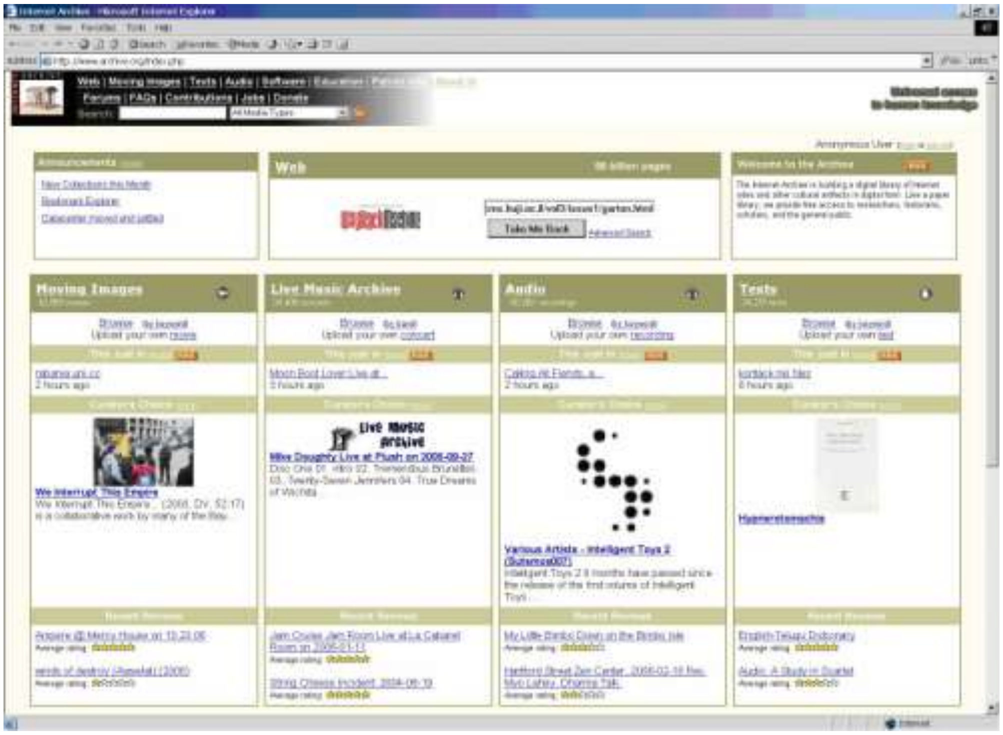
1) Access http://www.archive.org/index.php.


Web | Moving Images | Texts | Audio | Software | Education
Forums | FAQs | Contributions | Jobs | Donate
Search:
Web
Take Me Back
Advanced Search
Moving Images
Live Music Archive
Audio
Texts
Live Music Archive
Mike Doughty Live at Plush on 2000-09-27
We Interrupt This Empire
Various Artists - Intelligent Toys 2

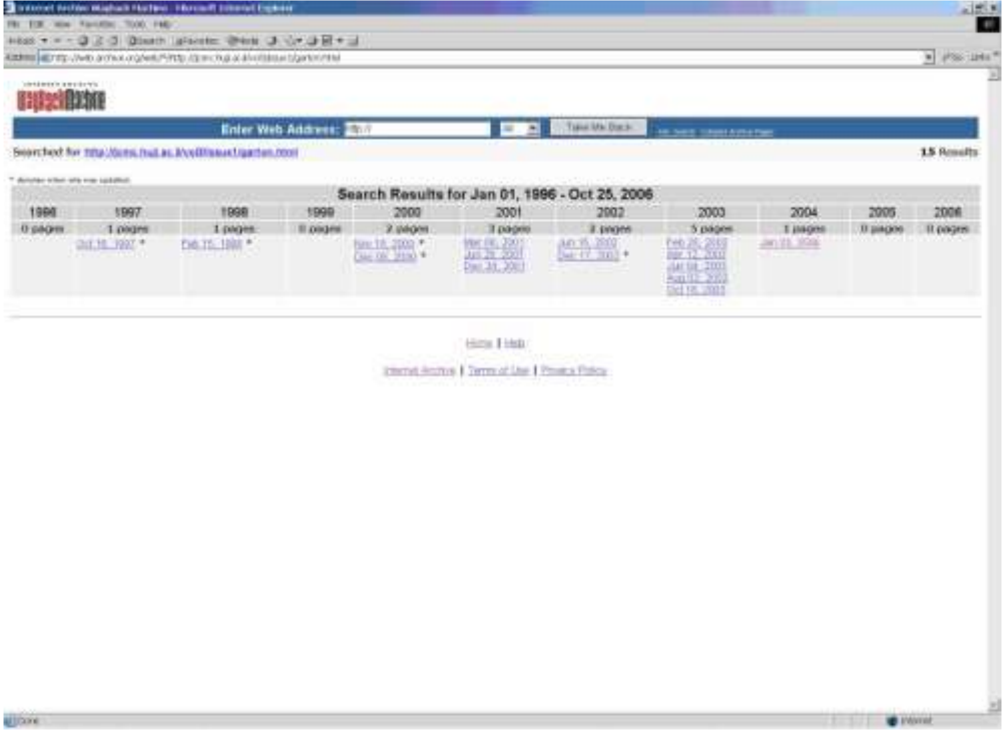
2) Type the URL address to retrieve in the search box.

(In this sample query, we retrieved an article titled "Studying online social networks" by Garton, L., Haythornthwaite, C., & Wellman, B., originally located at http://jcmc.huji.ac.il/vol3/issue1/garton.html)

3) Click on the "Take Me Back" button below search box.



4) The screen shot shows the result of the search.


Enter Web Address: http://
Take Me Back

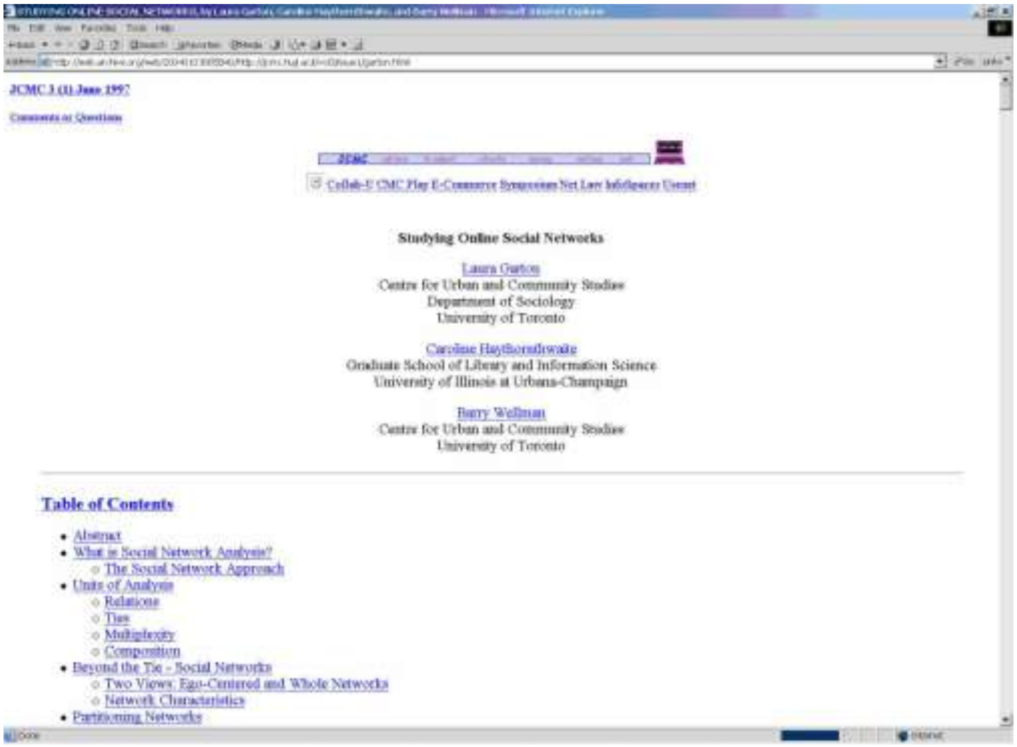
15 Results

Search Results for Jan 01, 1996 - Oct 25, 2006										
1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
0 pages	1 pages	1 pages	0 pages	2 pages	3 pages	2 pages	5 pages	1 pages	0 pages	0 pages
	Oct 18, 1997 *	Feb 15, 1998 *		Nov 18, 2000 * Dec 06, 2000 *	Mar 06, 2001 Jun 28, 2001 Dec 31, 2001	Jun 15, 2002 Dec 17, 2002 *	Feb 20, 2003 Mar 12, 2003 Jul 04, 2003 Aug 02, 2003 Oct 18, 2003	Jan 23, 2004		

Home | Help
Internet Archive | Terms of Use | Privacy Policy

5) Click one of hyperlinks in the table to retrieve the article.

6) Original article has been retrieved.


STUDYING ONLINE SOCIAL NETWORKS, by Laura Garton, Caroline Haythornthwaite, and Barry Wellman - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go Links
JCMC
Done
Internet

JCMC 3 (1) June 1997

Comments or Questions

Collab-U CMC Play E-Commerce Symposium Net Law InfoSpaces Usenet

Studying Online Social Networks

Laura Garton
Centre for Urban and Community Studies
Department of Sociology
University of Toronto

Caroline Haythornthwaite
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Barry Wellman
Centre for Urban and Community Studies
University of Toronto

Table of Contents