How Much of the Web Is Archived?

Scott G. Ainsworth, Ahmed AlSum, Hany SalahEldeen,
Michele C. Weigle, Michael L. Nelson
Old Dominion University Norfolk, VA, USA
{sainswor, aalsum, hany, mweigle, mln}@cs.odu.edu

ABSTRACT

The Memento Project's archive access additions to HTTP have enabled development of new web archive access user interfaces. After experiencing this *web time travel*, the inevitable question that comes to mind is "How much of the Web is archived?" This question is studied by approximating the Web via sampling URIs from DMOZ, Delicious, Bitly, and search engine indexes and measuring number of archive copies available in various public web archives. The results indicate that 35%–90% of URIs have at least one archived copy, 17%–49% have two to five copies, 1%–8% have six to ten copies, and 8%–63% at least ten copies. The number of URI copies varies as a function of time, but only 14.6–31.3% of URIs are archived more than once per month.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

General Terms

Design, Experimentation, Standardization

Keywords

Web Architecture, HTTP, Resource Versioning, Web Archiving, Temporal Applications, Digital Preservation

1. INTRODUCTION

With more and more of our business, academic, and cultural discourse contained primarily or exclusively on the Web, the problem of archiving the Web is receiving increased attention. The focal point of much of this attention is the Internet Archive's Wayback Machine, which began archiving the Web in 1996 and as of 2010 had over 1.5 billion unique URIs [10], making it the largest, longest-running and most well known publicly-available web archive. Recently, there has been a proliferation of new public web archives at universities, national libraries, and other organizations. These differ in scale, ingest models, collection development policies, and the software employed. All of this leads to the question "How much of the Web is archived?"

To address this question, we sampled URIs from four sources to estimate the percentage of archived URIs and the number and frequency of archived versions. From this, we extrapolate the percentage of the Surface Web that is archived.

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.

2. RELATED WORK

Although much has been written on the technical, social, legal, and political issues of web archiving, little research has focused on the archive coverage provided by existing archives. Day [2] surveyed a large number of archives while investigating the methods and issues associated with archiving but did not address coverage. Thelwall [13] touched on coverage when he addressed international bias in the Internet Archive, but did not directly address the percent of the Web that is covered. McCown and Nelson address coverage [8], but their research was limited to search engines caches.

Another open question is "how much coverage is required?" When Gomes, Freitas, et al. address the design of a national web archive, incompleteness is inherent in their compromise design [4]. Mason argues the Web and digital culture have changed our sense of permanence, which effected a change the collecting practices at the National Library of New Zealand [6]. Phillips systematically addresses archiving completeness and scope, but concludes that consensus has not been reached and that the Web's huge volume puts complete archiving out of reach [12].

Another web archiving issue that remained an obstacle until recently was the lack of standard APIs. The API with the most traction is Memento, which was proposed by Van de Sompel, Nelson, et al. [15]. Memento is an HTTP-based framework that bridges web archives and current resources. It provides a standards-based API for identifying and dereferencing archived resources using datetime negotiation. Each original resource, *URI-R*, has $i = 0..n$ archived representations, *URI-*M_i, that represent URI-R's states at times t_i. Using the Memento API, clients are able to request URI-M_i for a specified URI-R. Memento is now an IETF Internet Draft [14].

3. EXPERIMENT

From late November 2010 through early January 2011, we performed an experiment with the primary purpose of estimating the percentage of all publicly-visible URIs that have mementos available in public archives. The experiment

was accomplished in four parts: selecting URI sample sets, determining the current state of sample URIs, discovering mementos, and estimating sample URI age.

Our sample sets were selected from the Open Directory Project (DMOZ), Delicious, Bitly, and search engines. For practical reasons (e.g., search engine query limits and execution time) we used a sample size of 1,000 for all sample sets. Table 1 shows the mean number of mementos per URI-R, standard deviation, and standard error at a 95% confidence level for each sample set. Each sample set is listed twice: for all 1,000 URIs and for the subset of only those URIs that have at least one memento ($URI\text{-}M > 0$).

Table 1: Mementos per URI-R ($n = 1000$)

Collection	All			URI-M>0		
	Mean	SD	SE	Mean	SD	SE
DMOZ	56.85	119.35	7.40	62.68	123.86	7.68
Delicious	79.40	229.45	14.38	81.44	232.02	14.38
bitly	14.66	137.30	14.20	41.64	229.18	14.20
SE	5.40	22.55	1.40	6.99	25.40	1.57

Using the DMOZ as sample source has a long history [9, 5]. Although it is imperfect for many reasons (e.g., commercial bias), DMOZ was included for comparability with previous studies. It it is one of the oldest sources available which makes it a good source for URIs that may no longer exist.

Our DMOZ sampling differs from previous methods, such as Gulli and Signorini [5]. Instead of a snapshot in time, we used the entire available DMOZ history: 100 snapshots made from July 20, 2000 through October 3, 2010. This allowed old, non-existent URIs to be discovered and provided URI age estimation. After excluding non-HTTP and invalid URIs, DMOZ provided 9,415,486 unique URIs.

The second source for URIs is the social bookmarking site Delicious, which was started in 2003. It allows users to tag, save and share URIs. For the Delicious sample, we selected the first 1,000 URIs from the Delicious Recent random URI generator (`http://www.delicious.com/recent/?random`) on Nov. 22, 2010.

The Bitly project is a popular URI shortening service. It is Twitter's default URI shortener and has a significant user base. Bitly creates a short URI that redirects to a target URI when dereferenced. The Bitly URI consists of a 1–6 character, alphanumeric hash value appended to `http://bit.ly/` (e.g., `http://bit.ly/A`). For the Bitly sample, random hash values were created and dereferenced until 1,000 target URIs were discovered.

Search engines play an important role in web page discovery for most users of the Web. Previous studies have investigated the relationship between the Web as a whole and the subset indexed by search engines, which raised the need to select a sample from search engines indexes. Bar-Yossef and Gurevich have addressed search engine sampling in depth [1]. This experiment used their the phrase pool sampling method. The phrase pool was selected from the 5-grams in Google's N-gram data [3] and resulted in 1,176,470,663 queries. A random sample of these queries was used to obtain URIs, of which 1,000 were selected at random.

The current state of each sample URI was determined by the success (or failure) of dereferencing the URI and by the URI's indexing status in the Google, Bing, and Yahoo! search engines. Table 2 shows the results of dereferencing the URIs. Table 3 shows the number of indexed URIs per sample set. The Google web interface and API return significantly different results [7], so two Google statuses are shown: API-only and the union of the API and web interface.

Table 2: Sample URI Current HTTP Status

HTTP Status	DMOZ	Delicious	Bitly	SE
200	507	958	488	943
3xx⇒200	192	27	243	17
3xx⇒Other	50	1	36	3
4xx	135	8	197	16
5xx	4	3	6	0
Timeout	112	3	30	21

Table 3: Sample URI Search Engine Status

	DMOZ	Delicious	Bitly	SE
Bing	495	953	218	552
Yahoo	410	862	225	979
Google (API-only)	307	883	243	702
Google (API+web)	545	951	305	732

Memento discovery was conducted for the sample URIs using the Memento Project's proxies and aggregator [15].

Intuitively, the longer a URI is available on the Web, the greater the number of mementos we expect. Unfortunately, reliable creation dates are almost always unavailable [11]. To estimate the age of the URI, we use earliest creation time of the first memento, the first DMOZ archive containing the URI, and the time the URI was first added to Delicious.

4. RESULTS

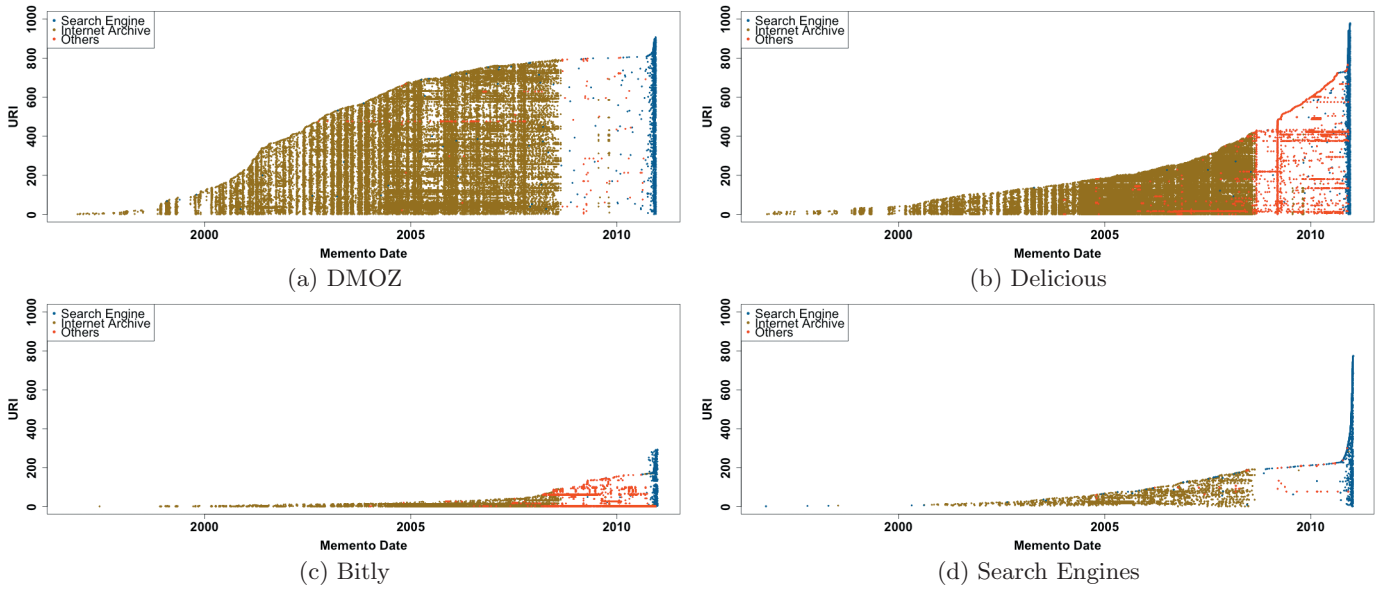
Figure 1 graphs the distribution of mementos over time. Three categories are shown: Internet Archive, search engine caches, and the other archives. The memento's date is on the x-axis and the URIs are on the y-axis. A dot represents a memento. Color indicates the source of the memento. Most of the mementos before 2008 are provided by the Internet Archive. Search engine caches provide very recent mementos. The archival rate for URIs with at least one memento is much higher for DMOZ and Delicious than for Bitly and search engines, which also differ considerably from each other. URIs from DMOZ and Delicious have a very high probability of being archived at least once while URIs from search engines have about 2/3 chance of being archived and Bitly URIs just under 1/3.

Table 4 summarizes the distribution of mementos. Two numbers stand out. First, the majority of Bitly URIs have no mementos. This matches the data in Table 3 and indicates poor coverage of the Bitly URIs by the search engines. Second, the majority of DMOZ URIs have more than 10 mementos. There are two likely causes for this: DMOZ is primary source for the Internet Archive and the DMOZ sample contains more old URIs than the other sources.

Table 4: Mementos Per URI.

Mementos per URI	DMOZ	Delicious	Bitly	SE
0 (Not archived)	93	25	648	225
1	46	79	100	336
2 – 5	142	491	171	320
6 – 10	85	35	17	35
More than 10	634	370	64	84

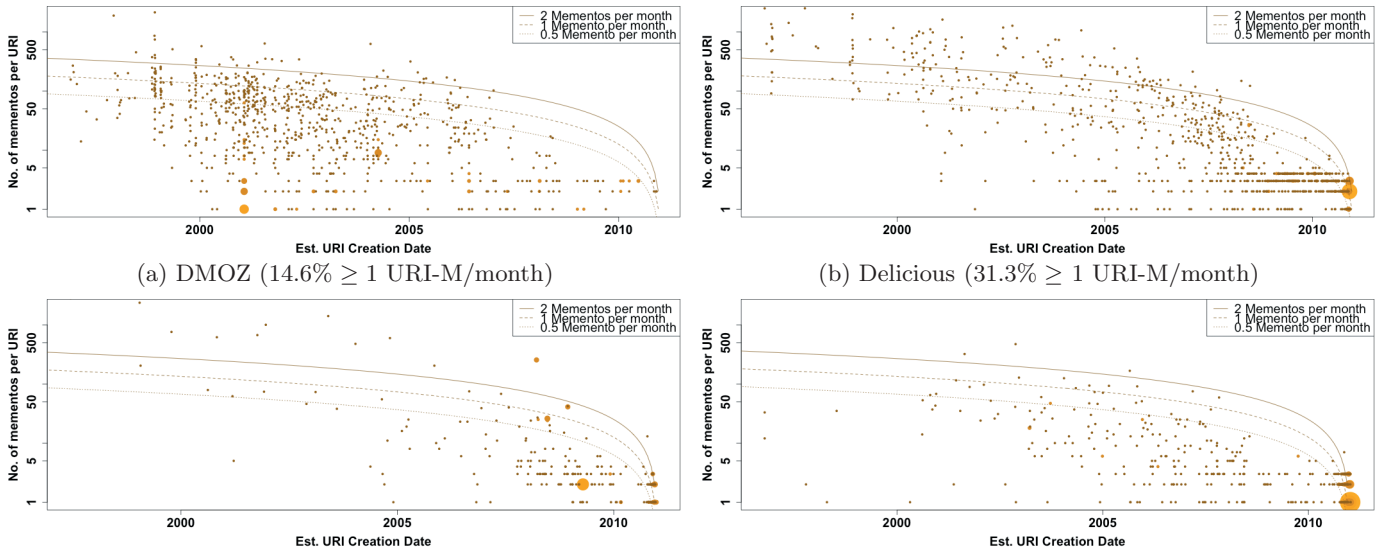
Figure 2 shows the density of mementos by estimated URI creation date, which is on the x-axis. The y-axis is the number of mementos for this URI. Density guidelines are shown for 0.5, 1, and 2 mementos per month.



(a) DMOZ (b) Delicious

(c) Bitly (d) Search Engines

Figure 1: Memento Distribution, ordered by the first observation date.



(a) DMOZ (14.6% $\geq$ 1 URI-M/month) (b) Delicious (31.3% $\geq$ 1 URI-M/month)

(c) Bitly (14.7% $\geq$ 1 URI-M/month) (d) Search Engines (28.7% $\geq$ 1 URI-M/month)

Figure 2: The relation between the Mementos' density and the URI's age in months.

Table 5 shows retrieved memento statistics. For each sample source, we show the number of mementos (URI-M) discovered, the number of URIs (URI-R) that have mementos, the mean number of mementos, and standard deviation.

5. ANALYSIS

Our research indicates that URI source is a major driver of archival status. Common to DMOZ and Delicious is that URIs are added by people. Search engine sample URIs are discovered through crawling, a less selective process depending on search engine crawl heuristics instead of direct human activity. Bitly is more of a mystery (our research did not look into how Bitly URIs are used); however, the low archival rate leads us to think that many private, standalone, or temporary URIs were discovered by our selection algorithm.

The number of backlinks is an indication of the popularity of a page. The number of backlinks for each URI was calculated using Yahoo Boss APIs and we used Kendall's Tau ($p \leq 0.05$) to test the correlation between the number of backlinks and number of mementos. There is a weak positive relationship between the number of backlinks and number of mementos. The correlations are: DMOZ 0.389, Delicious 0.311, Bitly 0.631, and search engines 0.249. This positive correlation and the archival rate differences suggests that archival rate is influenced by URI popularity.

We also included a variety of archives in order to ascertain differences between them. For comparison, the archives are divided into three groups: the Internet Archive (IA), search

Table 5: Coverage by Archive.

Archive	DMOZ				Delicious				Bitly				Search Engines			
	#URI-M	#URI-R	Mean	SD	#URI-M	#URI-R	Mean	SD	#URI-M	#URI-R	Mean	SD	#URI-M	#URI-R	Mean	SD
Internet Archive	55293	783	70.62	130	74809	408	183.36	325	8947	70	127.81	406	4067	170	23.92	49
Google	523	523	1	0	897	897	1	0	253	253	1	0	486	486	1	0
Bing	427	427	1	0	786	786	1	0	204	204	1	0	515	515	1	0
Yahoo	418	418	1	0	479	479	1	0	87	87	1	0	229	229	1	0
Diigo	36	36	1	0	354	354	1	0	61	61	1	0	10	10	1	0
Archive-It	92	4	23	41	500	38	13.16	30	75	13	5.77	8	49	12	4	5
National Archives (UK)	25	8	3.125	3	521	102	5.11	10	531	12	44.25	145	1	1	1	0
NARA	5	5	1	0	31	19	1.63	1	10	2	5	6	4	2	2	0
UK Web Archive	8	5	1.6	1	391	38	10.29	16	2892	32	90.38	187	9	3	3	3
WebCite	26	5	5.2	8	594	57	10.42	49	989	58	17.05	82	–	–	–	–
ArchiefWeb	–	–	–	–	22	3	7.33	11	609	1	609	0	–	–	–	–
CDLIB	–	–	–	–	20	5	4	4	–	–	–	–	–	–	–	–

engines, and all others (which tend to be specialized). These groups are compared on URI coverage, depth, and age.

Coverage is the number of URIs in the archive. The Internet Archive and search engines have comparable URI coverage: 35–90% have at least 1 memento, 17–49% have 2–5, 1–8% have 6–10, and 8–63% more than 10. The other archives are specialized and cover only a minor fraction of the Web.

The Internet Archive also has the greatest depth, which is the number of mementos per URI. This is probably because it is the oldest. The search engines have the least depth—1 memento—by design, and age is irrelevant. Most other archives are somewhere in the middle and only cover the past few years.

6. CONCLUSIONS

Although our research shows 35–90% of public URIs have at least one memento, coverage is inconsistent and appears dependent on several factors. Human desire for URI publicity appears to be a major factor as shown by the relatively high DMOZ and Delicious archival rates. Search engine discoverability is the next most important factor followed by explicit archiving. The best overall coverage is provided by the Internet Archive. The search engines follow, but only for very recent mementos. The specialized archives provide good coverage for the URIs they cover (but only for the URIs they cover).

Future work will include study of the relationship between the rate of change of the URI and the rate of the archiving process. This work has been done on a general sample of URIs. In future work, archived URIs will be studied based on specific languages beyond English.

7. ACKNOWLEDGMENTS

This work is supported in part by the Library of Congress. We thank Herbert Van de Sompel and Robert Sanderson of Los Alamos National Laboratory and Kris Carpenter Negulescu and Bradley Tofel of the Internet Archive for their explanations and positive comments. Bar-Yossef graciously provided us with source code for search engine sampling.

8. REFERENCES

[1] Ziv Bar-Yossef and Maxim Gurevich. Random sampling from a search engine's index. *J. ACM*, 55(5), 2008.

[2] Michael Day. Preserving the fabric of our lives: A survey of web preservation initiatives. In *ECDL'03*, 2003.

[3] Alex Franz and Thorsten Brants. All our n-gram are belong to you. `http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html`, August 2006. Accessed January 15, 2011.

[4] Daniel Gomes, Sérgio Freitas, and Mário Silva. Design and selection criteria for a national web archive. In *ECDL'06*, 2006.

[5] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *WWW'05*, 2005.

[6] Ingrid Mason. Virtual preservation: How has digital culture influenced our ideas about permanence? changing practice in a national legal deposit library. *Library Trends*, 56(1), Summer 2007.

[7] Frank McCown and Michael L. Nelson. Agreeing to disagree: search engines and their public interfaces. In *JCDL'07*, 2007.

[8] Frank McCown and Michael L. Nelson. Characterization of search engine caches. In *Proceedings of IS&T Archiving 2007*, May 2007.

[9] G. Monroe, J. French, and A. Powell. Obtaining language models of web collections using query-based sampling techniques. *HICSS'02*, 3, 2002.

[10] Kris Carpenter Negulescu. Web archiving @ the Internet Archive. `http://www.digitalpreservation.gov/news/events/ndiipp_meetings/ndiipp10/docs/July21/session09/NDIIPP072110FinalIA.ppt`, 2010.

[11] Michael L. Nelson. Web Science and Digital Libraries Research Group: 2010-11-05: Memento-Datetime is not Last-Modified, 2010. `http://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not-last.html`.

[12] Margaret E. Philips. What should we preserve? the question for heritage libraries in a digital world. *Library Trends*, 54(1), Summer 2005.

[13] Mike Thelwall and Liwen Vaughan. A fair history of the Web? examining country balance in the Internet Archive. *Library & Information Science Research*, 26(2), 2004.

[14] Herbert Van de Sompel, Michael Nelson, and Robert Sanderson. HTTP framework for time-based access to resource states — Memento, November 2010. `http://datatracker.ietf.org/doc/draft-vandesompel-memento/`.

[15] Herbert Van de Sompel, Michael L. Nelson, Robert Sanderson, Lyudmila L. Balakireva, Scott Ainsworth, and Harihar Shankar. Memento: Time travel for the web. Technical Report arXiv:0911.1112, 2009.