How Much of the Web is Archived?*

Scott Ainsworth, Ahmed AlSum, Hany SalahEldeen,
Michele C. Weigle, Michael L. Nelson

Old Dominion University, USA
{sainswor, aalsum, hany, mweigle, mln}@cs.odu.edu

JCDL 2011, Ottawa, Canada

*This work supported in part by the Library of Congress and NSF IIS-1009392.

How Much of the Web is Archived?*

Scott Ainsworth, Ahmed AlSum, Hany SalahEldeen,
Michele C. Weigle, Michael L. Nelson

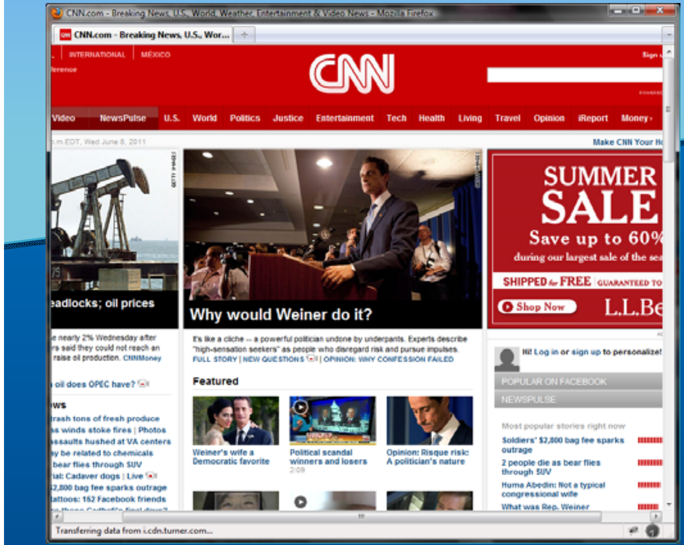
Old Dominion University, USA
{sainswor, aalsum, hany, mweigle, mln}@cs.odu.edu

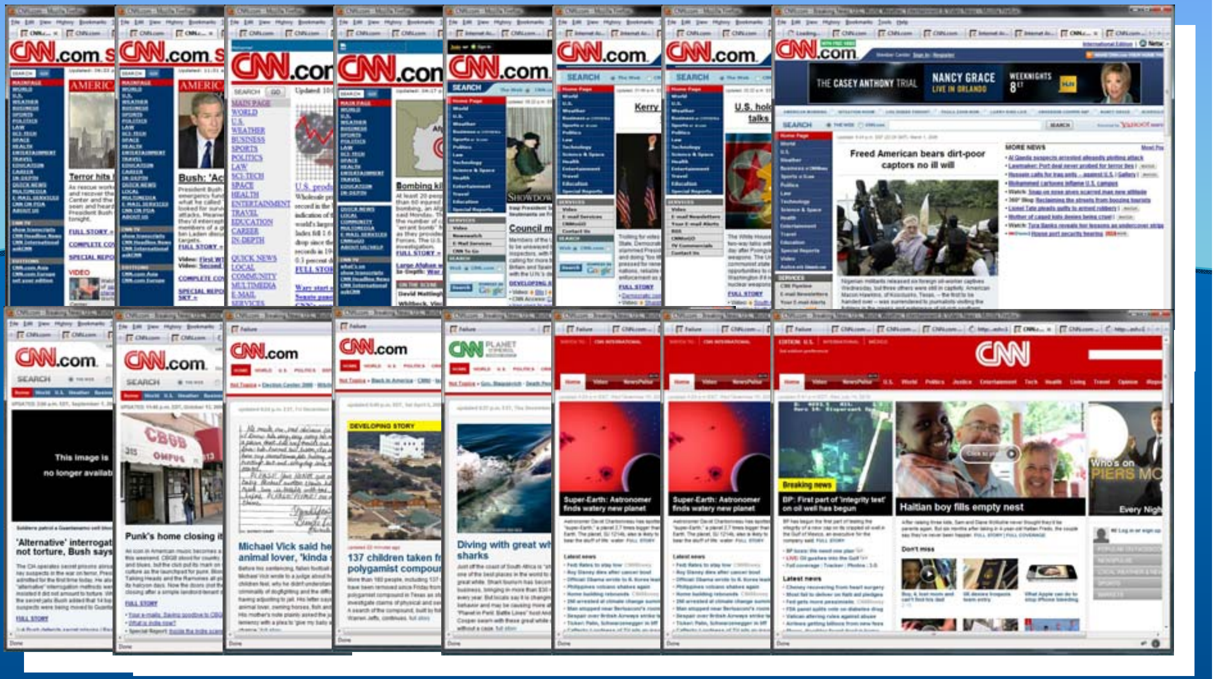
JCDL 2011, Ottawa, Canada

*This work supported in part by the Library of Congress and NSF IIS-1009392.

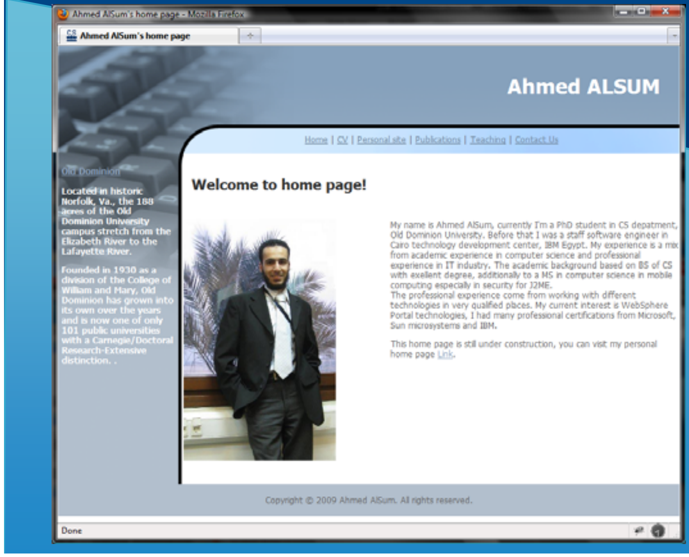
How Many Archive Copies?

http://www.cnn.com


CNN.com - Breaking News, U.S., World, Weather, Entertainment & Video News - Mozilla Firefox
CNN
Video
NewsPulse
U.S.
World
Politics
Justice
Entertainment
Tech
Health
Living
Travel
Opinion
iReport
Money
Why would Weiner do it?
SUMMER SALE
Save up to 60%
SHIPPED for FREE
Shop Now
L.L.Bean
Featured
Weiner's wife a Democratic favorite
Political scandal winners and losers
Opinion: Risque risk: A politician's nature
POPULAR ON FACEBOOK
NEWSPULSE
Transferring data from i.cdn.turner.com...


CNN.com
Terror hits
Bush: 'Ac
Bombing ki
Council m
Freed American bears dirt-poor captors no ill will
THE CASEY ANTHONY TRIAL
NANCY GRACE LIVE IN ORLANDO
'Alternative' interrogat not torture, Bush says
Punk's home closing
Michael Vick said he animal lover, 'kinda
137 children taken fr polygamist compou
Diving with great wh sharks
Super-Earth: Astronomer finds watery new planet
BP: First part of 'integrity test' on oil well has begun
Haitian boy fills empty nest
Who's on PIERS MO

http://cs.odu.edu/~aalsum



Ahmed ALSUM

Home | CV | Personal site | Publications | Teaching | Contact Us

Old Dominion

Located in historic Norfolk, Va., the 188 acres of the Old Dominion University campus stretch from the Elizabeth River to the Lafayette River.

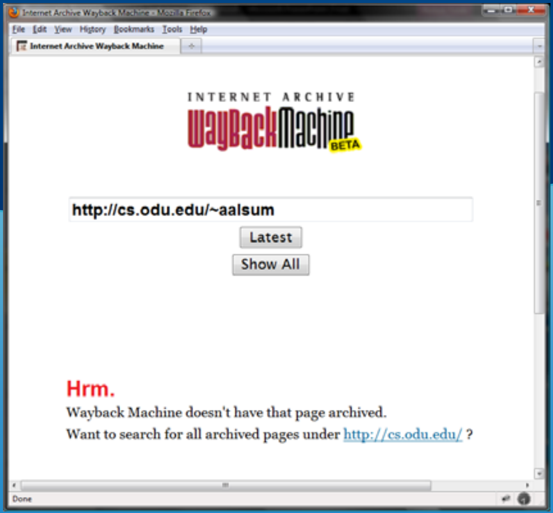
Founded in 1930 as a division of the College of William and Mary, Old Dominion has grown into its own over the years and is now one of only 101 public universities with a Carnegie/Doctoral Research-Extensive distinction. .

Welcome to home page!

My name is Ahmed AlSum, currently I'm a PhD student in CS depatment, Old Dominion University. Before that I was a staff software engineer in Cairo technology development center, IBM Egypt. My experience is a mix from academic experience in computer science and professional experience in IT industry. The academic background based on BS of CS with exellent degree, additionally to a MS in computer science in mobile computing especially in security for J2ME.
The professional experience come from working with different technologies in very qualified places. My current interest is WebSphere Portal technologies, I had many professional certifications from Microsoft, Sun microsystems and IBM.

This home page is still under construction, you can visit my personal home page Link.

Copyright © 2009 Ahmed AlSum. All rights reserved.


Internet Archive Wayback Machine - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Internet Archive Wayback Machine
INTERNET ARCHIVE
WaybackMachine BETA
Latest
Show All
Done

http://cs.odu.edu/~aalsum

Hrm.

Wayback Machine doesn't have that page archived.

Want to search for all archived pages under http://cs.odu.edu/ ?

Research Question

How Much of the Web is Archived?

Experiment

Sampling Techniques

- 4 Sample sets – 1000 URIs each

dmoz open directory project

delicious

bitly

Sampling Techniques

DMOZ

dmoz open directory project

- **All the available DMOZ history:**
  100 snapshots made from July 20, 2000 through October 3, 2010
- **Remove duplicates, non-HTTP, invalid URIS**
- **Randomly selected 1000 URIs**

HTTP Status	#
200	507
3xx→200	192
3xx→ Others	50
4xx	135
5xx	4
Timeout	112

Indexed by:	bing	YAHOO!	Google API	Google API + Web
	49%	41%	30%	54%

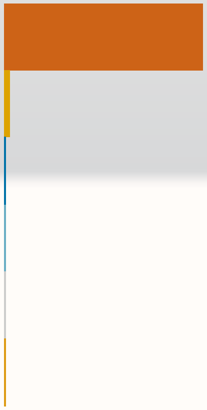
Sampling Techniques

Delicious

delicious

- **We used Delicious Recent random URI generator** (http://www.delicious.com/recent/?random=1) **on Nov. 22, 2010 to get 1000 URIs.**

HTTP Status	#
200	958
3xx→200	27
3xx→ Others	1
4xx	8
5xx	3
Timeout	3



Indexed by:	bing	YAHOO!	Google API	Google API + Web
	95%	86%	88%	95%

Sampling Techniques

Bitly

bitly

- Bitly URI consists of a 1–6 character, alphanumeric hash value appended to http://bit.ly/
- Random hash values were created and dereferenced until 1,000 target URIs were discovered

HTTP Status	#
200	488
3xx→200	243
3xx→ Others	36
4xx	197
5xx	6
Timeout	30

Indexed by:	bing	YAHOO!	Google API	Google API + Web
	21%	22%	24%	30%

Sampling Techniques

Search Engine

- **Random Sample from Search Engine technique** [Bar-Yossef 2008]
- **The phrase pool was selected from the 5-grams in Google's N-gram data.**
- **A random sample of these queries was used to obtain URIs, of which 1,000 were selected at random.**

HTTP Status	#
200	943
3xx→200	17
3xx→ Others	3
4xx	16
5xx	0
Timeout	21

Indexed by:	bing	YAHOO!	Google API	Google API + Web
	55%	98%	70%	73%

Experiment

- **For each sample set, we used Memento Aggregator to get all the possible archived copies (Mementos).**
- **For each URI, Memento Aggregator responded with TimeMap for this URI.**

Example

<http://memento.waybackmachine.org/memento/20010819194233/http://jcdl2002.org>;rel="first memento";datetime="Sun, 19 Aug 2001 19:42:33 GMT“, <http://memento.waybackmachine.org/memento/20011216220248/http://jcdl2002.org>; rel="memento"; datetime="Sun, 16 Dec 2001 22:02:48 GMT",

Results

#Mementos	dmoz	delicious	bitly	SE YAHOO!
0 (Not archived)	10%	3%	65%	22%
1	4%	8%	10%	34%
2–5	14%	48%	17%	32%
6–10	9%	4%	2%	4%
More than 10	63%	37%	6%	8%

Archive categories

We have 3 categories of archives

- Internet Archive (classic interface)

INTERNET ARCHIVE

- Search engine

bing YAHOO! Google

- Other archives

BRITISH LIBRARY

ARCHIVE-IT

WebCite

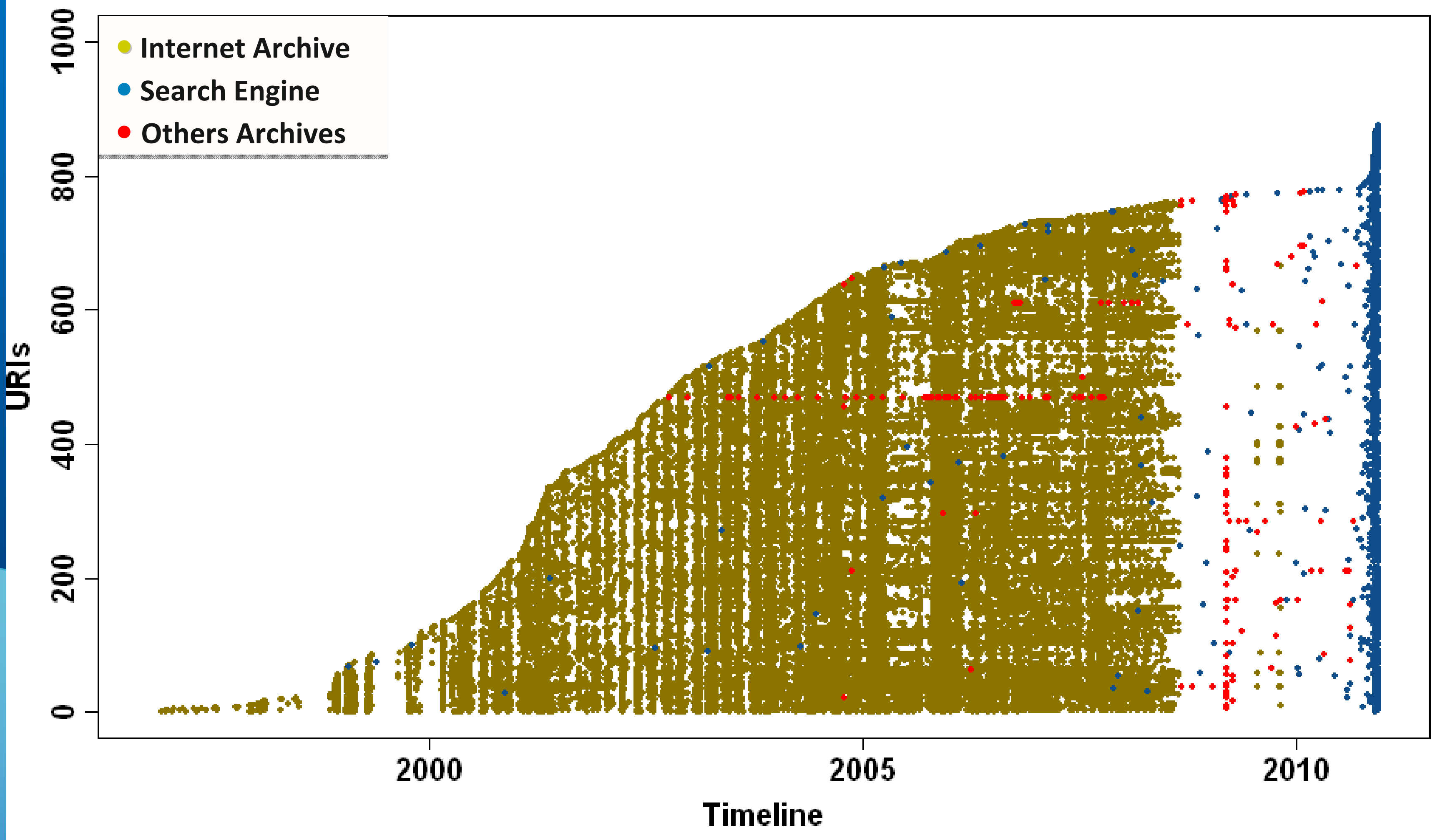
archiefweb.eu

University of California CDL California Digital Library

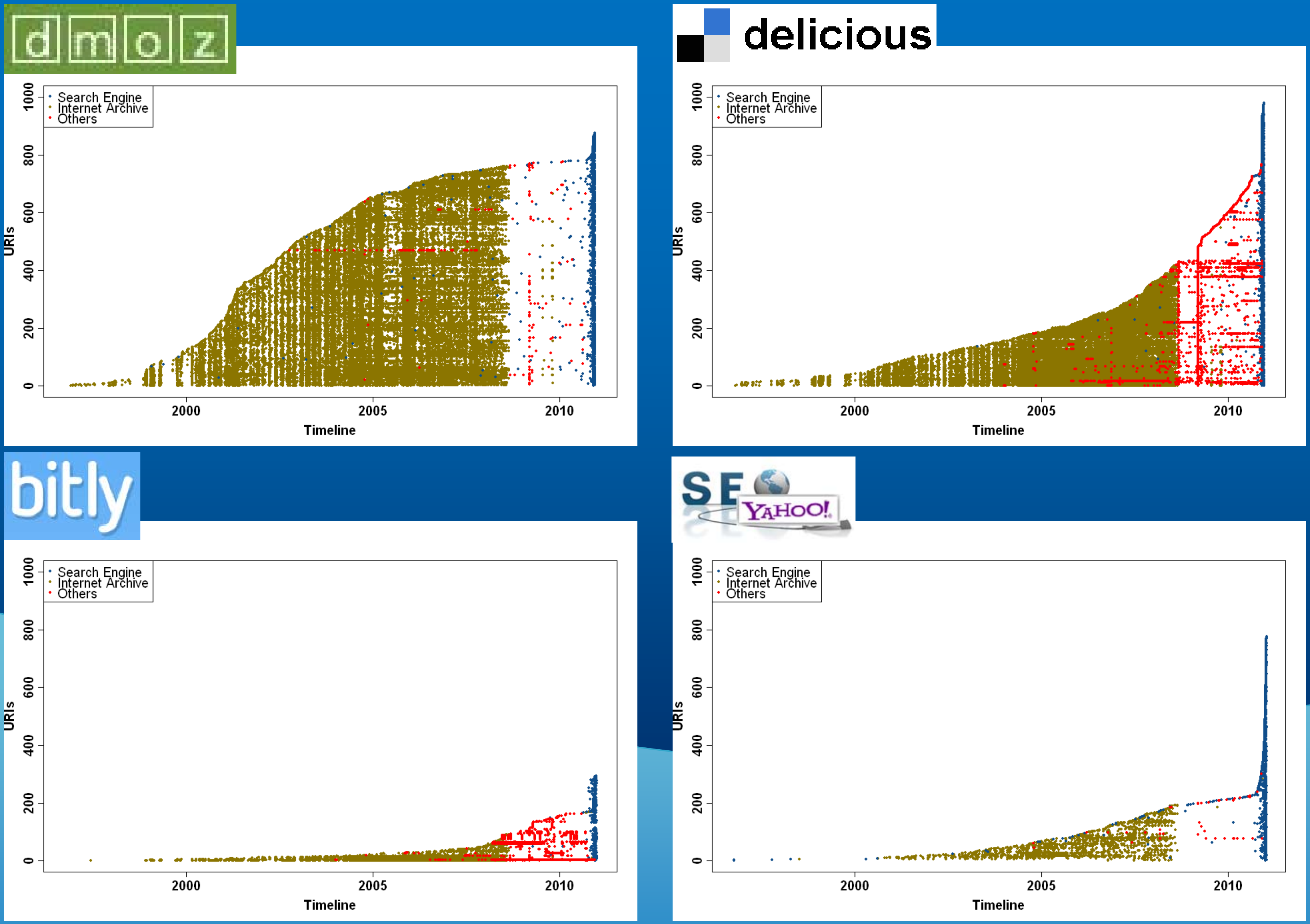
The National Archives UK

NATIONAL ARCHIVES US

diigo V5.0 beta

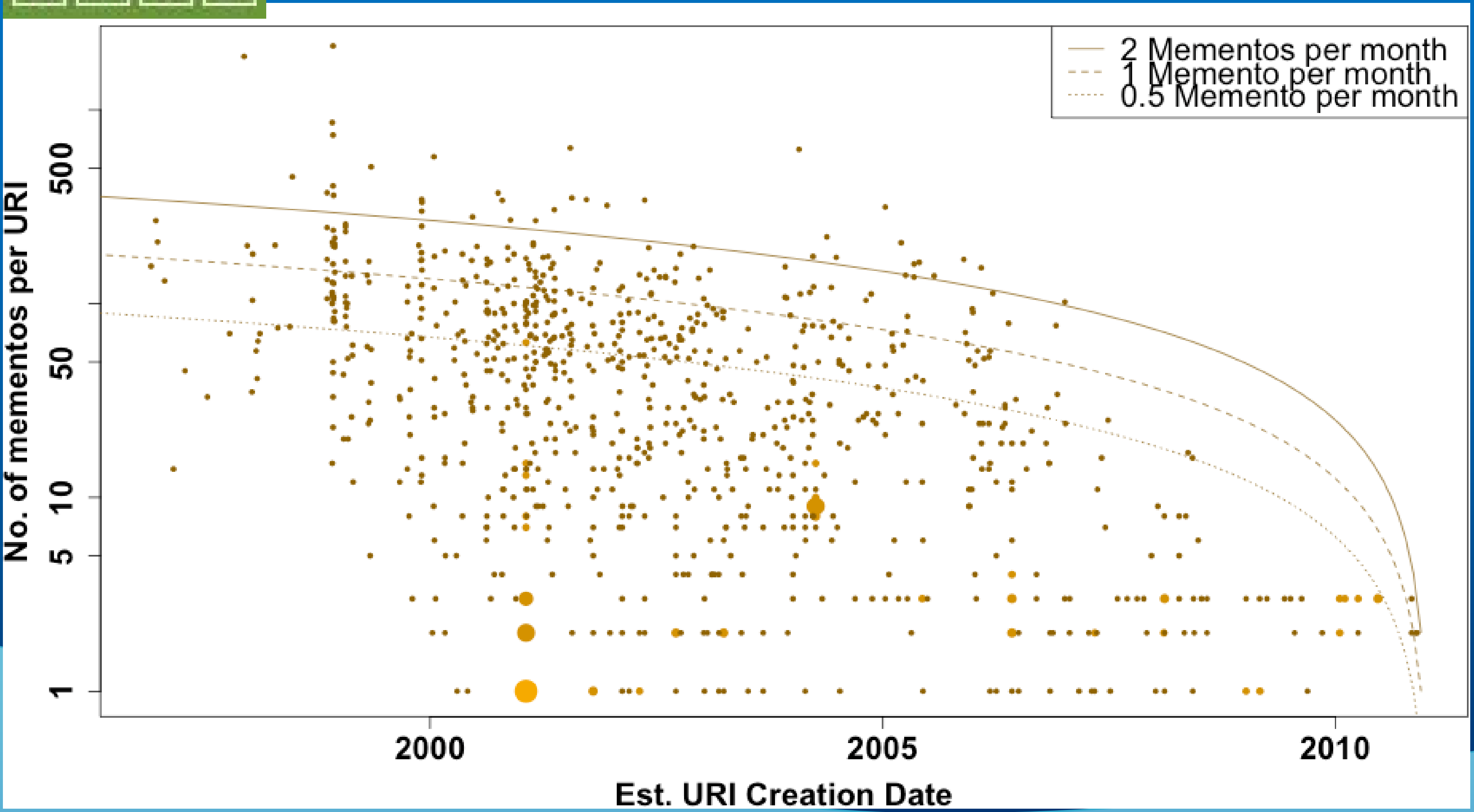


Memento Distribution, ordered by the first observation date.



Memento Distribution, ordered by the first observation date.

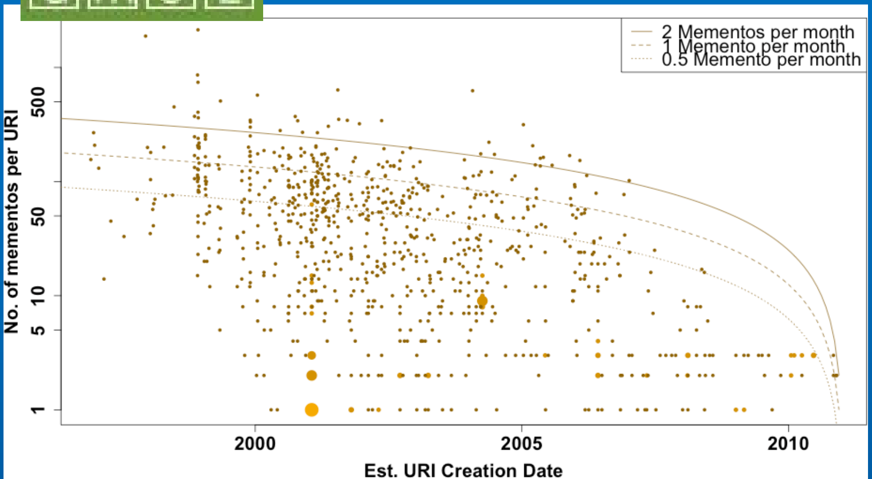
dmoz



(14.6% archived >= once per month)

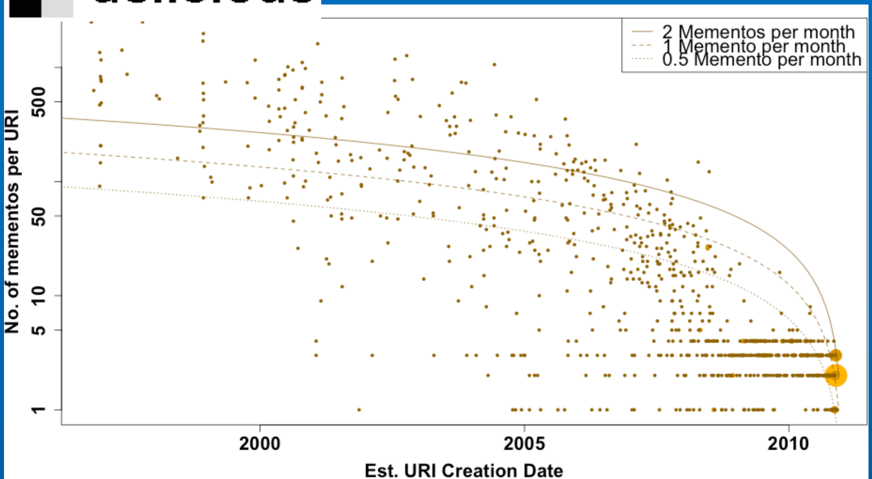
How often are URIs archived?

dmoz



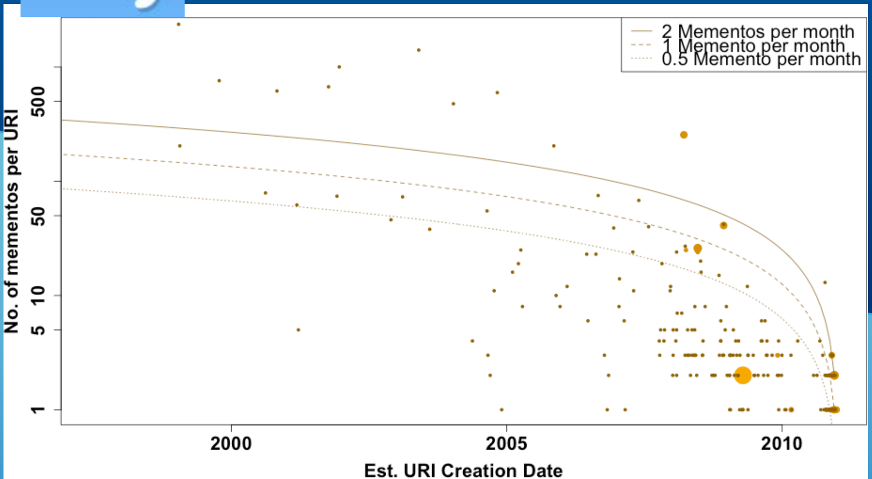
(14.6% archived >= once per month)

delicious



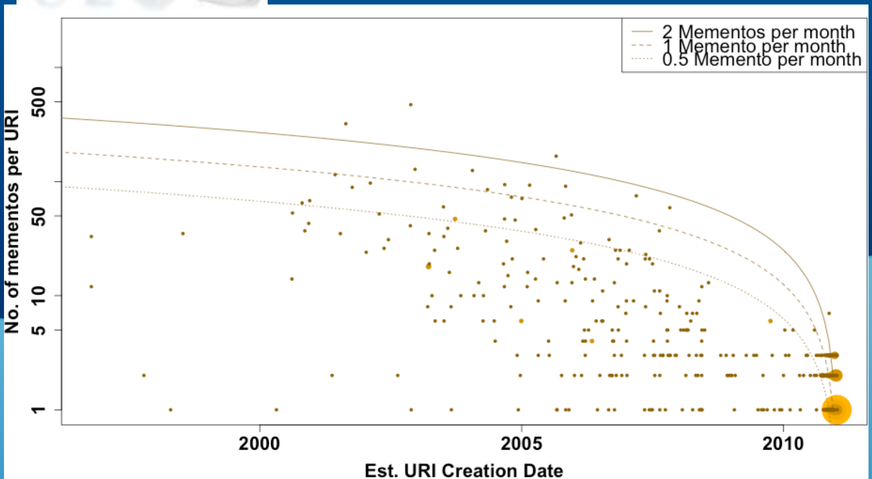
(31% archived >= once per month)

bitly



(14.7% archived >= once per month)

SE YAHOO!



(28.7% archived >= once per month)

How often are URIs archived?

Analysis

- What affects the archival rate?
  - – URI source (More popular – More archival)
  - – BackLinks (Weak positive relationship)
- Archive Coverage
  - – Internet Archive (Best)
  - – Search Engine (Good), but only one copy.
  - – The other archives cover only a small fraction with recent copies.

Conclusion

How much of the Web is Archived?

- Tell me what is your URI source!!

	Including SE cache	Excluding SE Cache
dmoz	90%	79%
delicious	97%	68%
bitly	35%	16%
SE YAHOO!	88%	19%

References

- Ziv Bar-Yossef and Maxim Gurevich. Random sampling from a search engine's index. *J. ACM*, 55(5), 2008.
- Frank McCown and Michael L. Nelson. Characterization of search engine caches. In *Proceedings of IS&T Archiving 2007*, May 2007.
- 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-vandesompelmemento/.

Thank You

Contact:

Ahmed AlSum

aalsum@cs.odu.edu