Old Dominion University ODU Digital Commons

Computer Science Presentations

Computer Science

9-6-2013

Who Will Archive the Archives? Thoughts About the Future of Web Archiving

Michael L. Nelson Old Dominion University, mnelson@odu.edu

Scott G. Ainsworth Old Dominion University

Ahmed Alsum Old Dominion University

Justin F. Brunelle Old Dominion University

Mat Kelly Old Dominion University

See next page for additional authors

Follow this and additional works at: https://digitalcommons.odu.edu/ computerscience presentations



Part of the Archival Science Commons

Recommended Citation

Nelson, Michael L.; Ainsworth, Scott G.; Alsum, Ahmed; Brunelle, Justin F.; Kelly, Mat; SalahEldeen, Hany; Weigle, Michele C.; Sanderson, Robert; and de Sompel, Herbert Van, "Who Will Archive the Archives? Thoughts About the Future of Web Archiving" (2013). Computer Science Presentations. 15.

https://digitalcommons.odu.edu/computerscience_presentations/15

This Book is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Presentations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

Authors Michael L. Nelson, Scott G. Ainsworth, Ahmed Alsum, Justin F. Brunelle, Mat Kelly, Hany SalahEldeen, Michael C. Weigle, Robert Sanderson, and Herbert Van de Sompel	





Who Will Archive the Archives? Thoughts About the Future of Web Archiving

Michael L. Nelson

Old Dominion University

with:

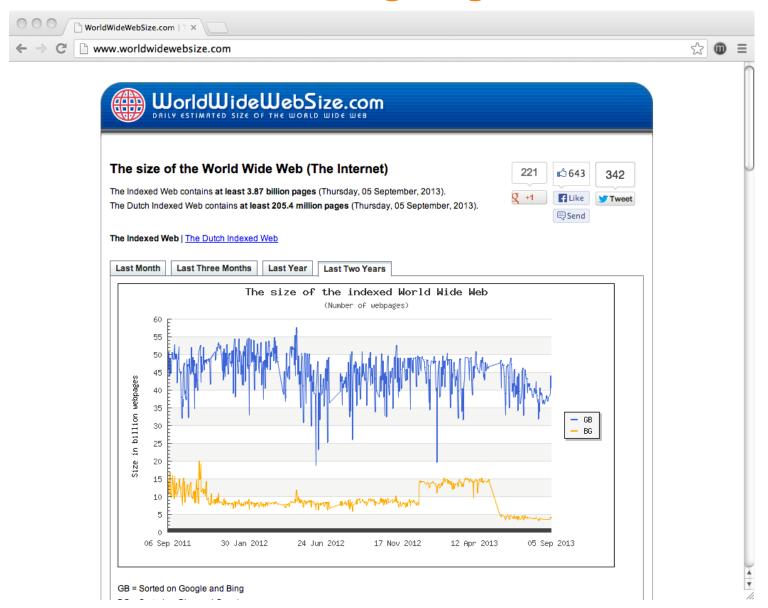
Old Dominion University: Scott G. Ainsworth, Ahmed AlSum, Justin F. Brunelle, Mat Kelly, Hany SalahEldeen, Michele C. Weigle

Los Alamos National Laboratory: Robert Sanderson, Herbert Van de Sompel





Web Archiving: Big Data?







Two Common Misconceptions About Web Archiving

- Prior = old = obsolete = stale = bad
 - who cares, not an interesting problem
- The Internet Archive has every copy of everything that has ever existed
 - who cares, problem solved





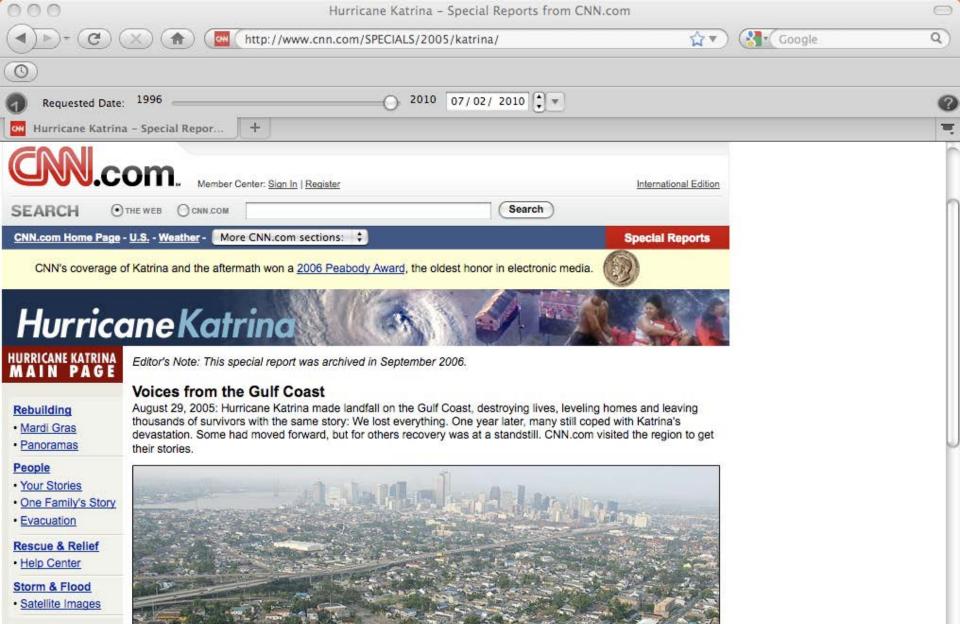
Why Care About The Past?

From an anonymous WWW 2010 reviewer about our Memento paper (*emphasis* mine):

"Is there any statistics to show that many or a good number of Web users would like to get obsolete data or resources?"

one answer: replay of contemporary pages >> summary pages

http://www.slideshare.net/phonedude/why-careaboutthepast http://www.nytimes.com/2013/06/19/books/seven-american-deaths-and-disasters-transcribes-the-news.html

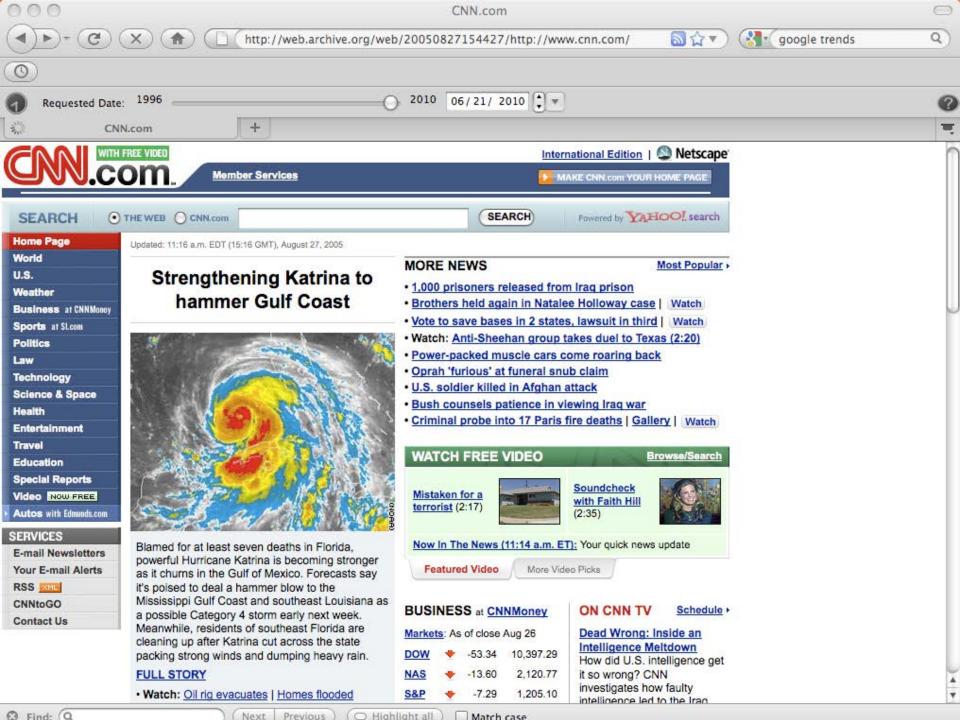


Hurricane Season Special Report

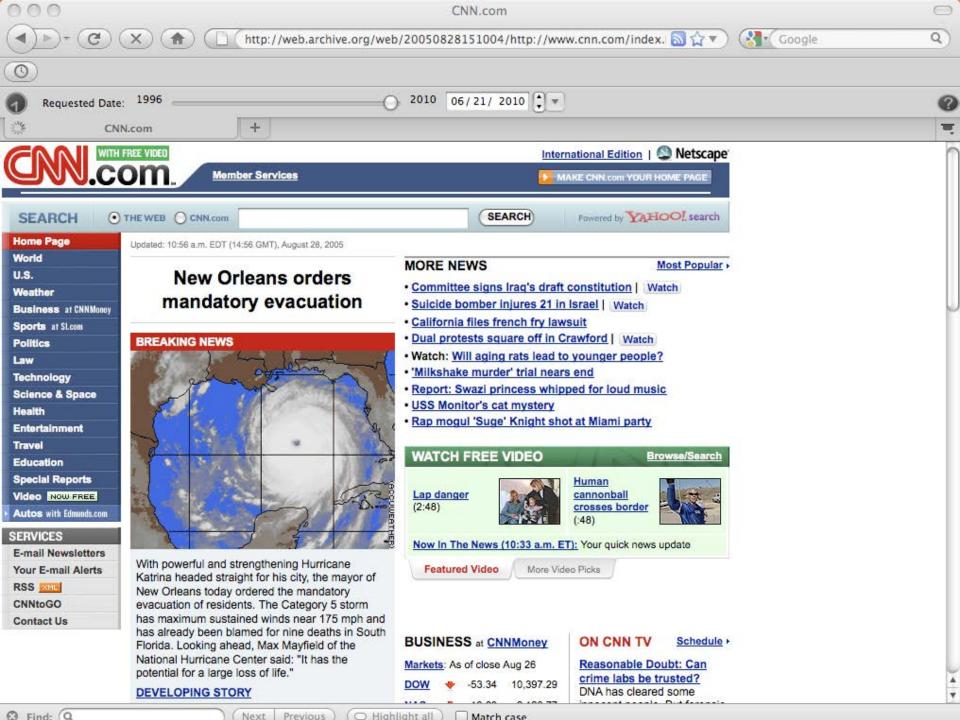


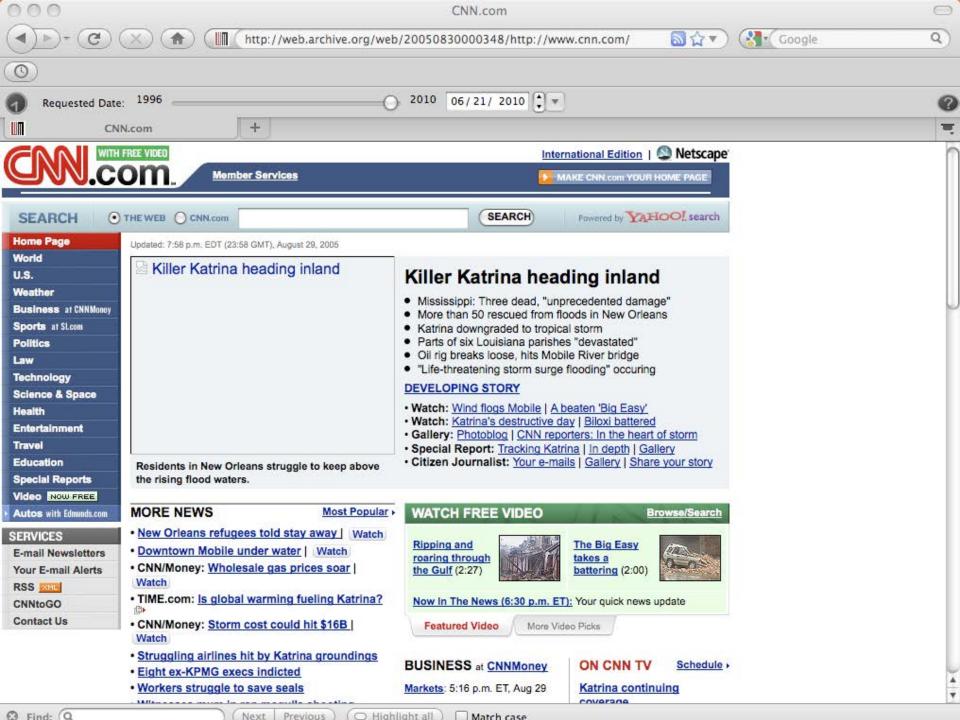


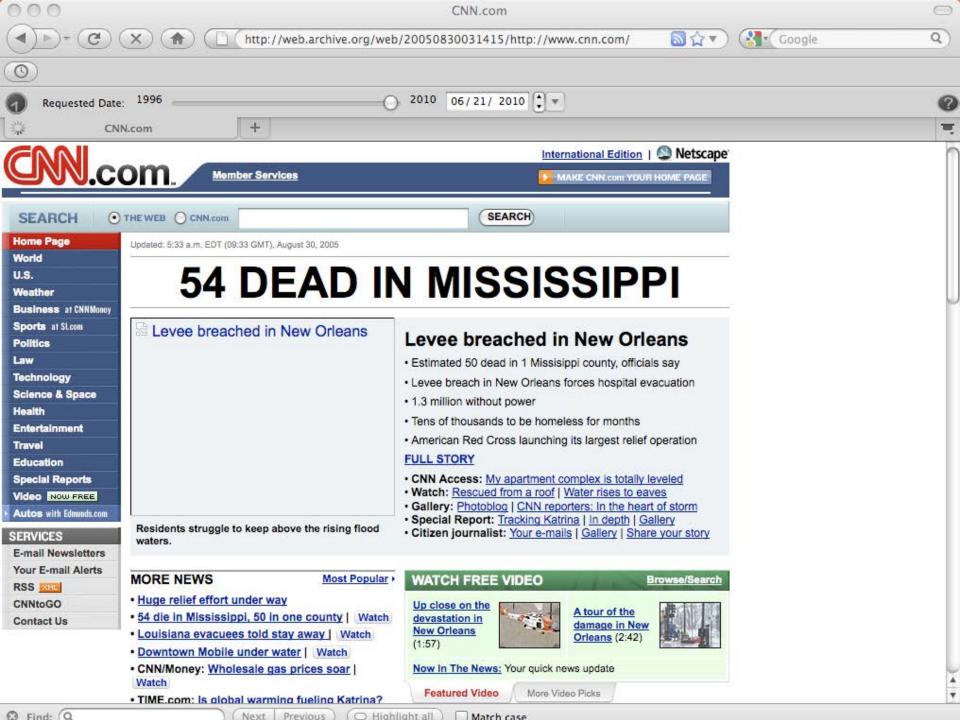
VS.

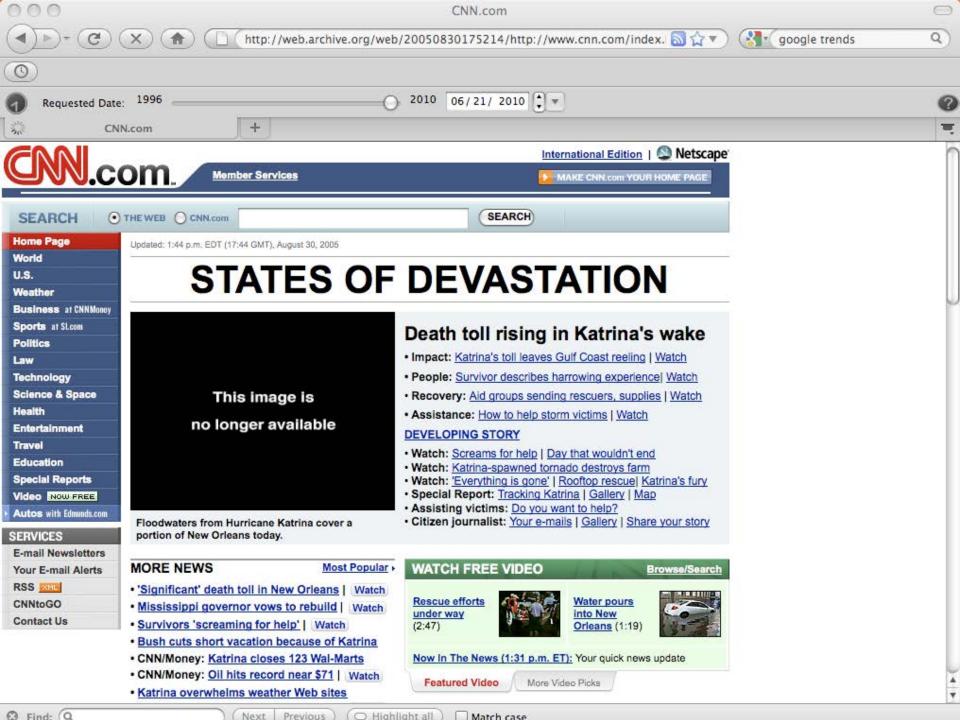








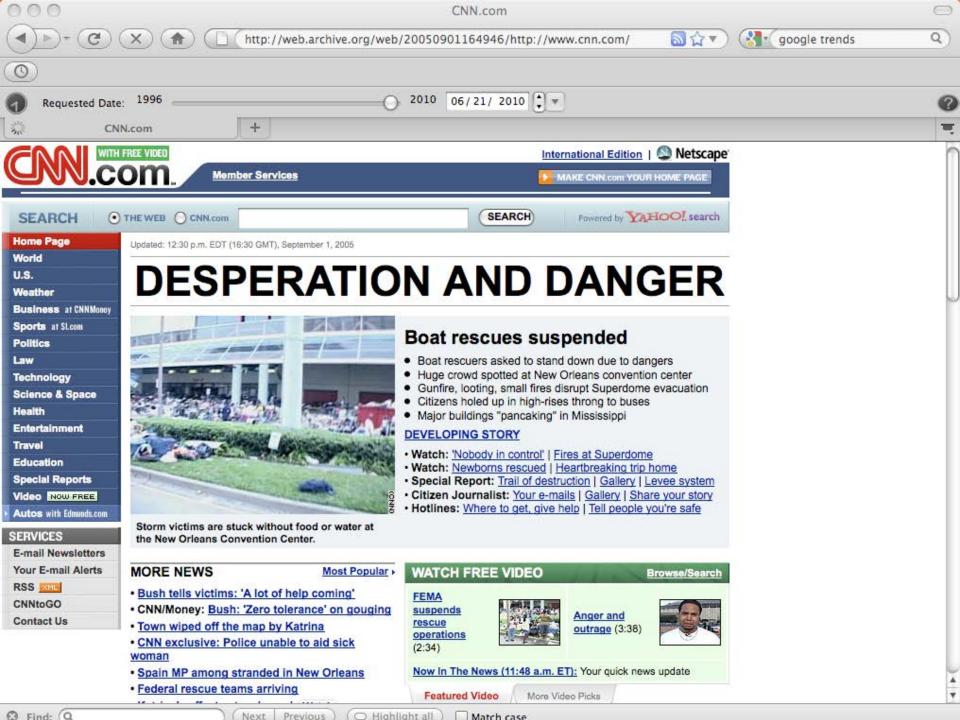




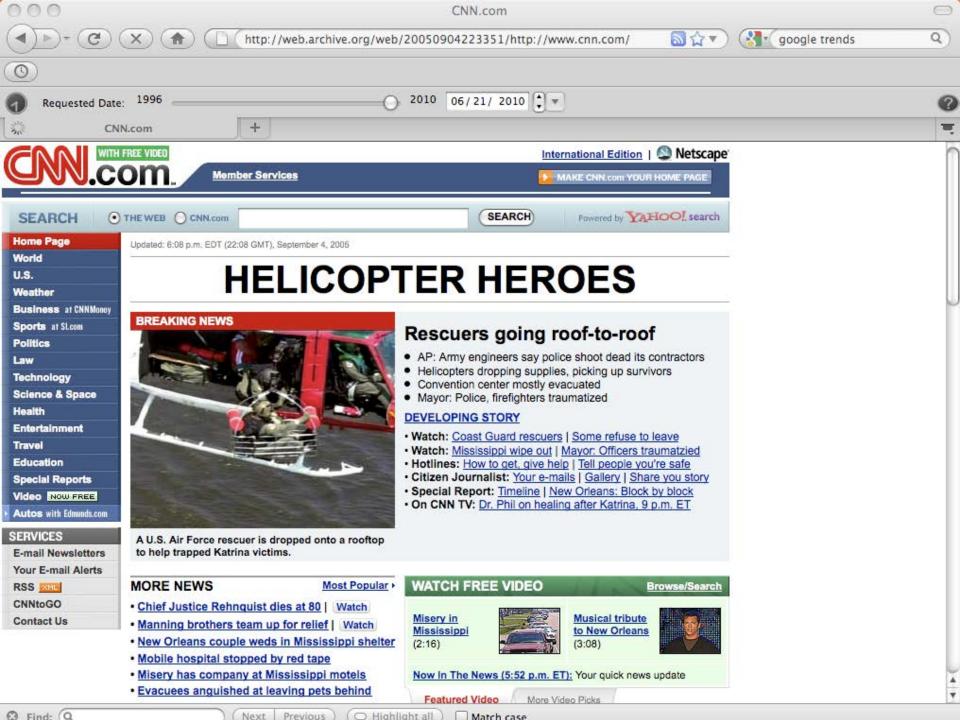








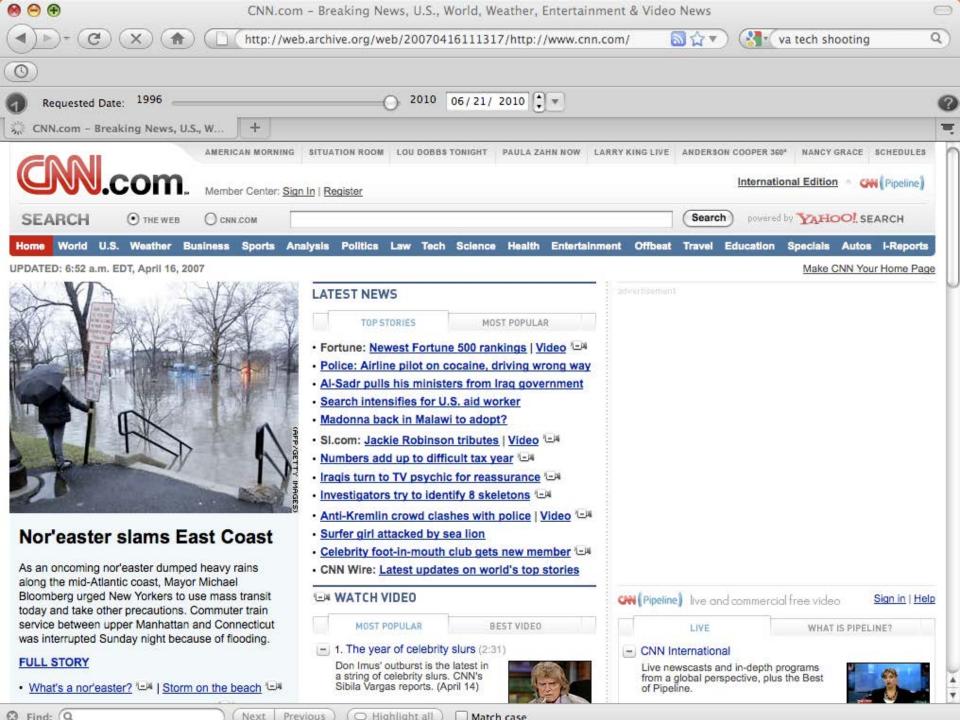




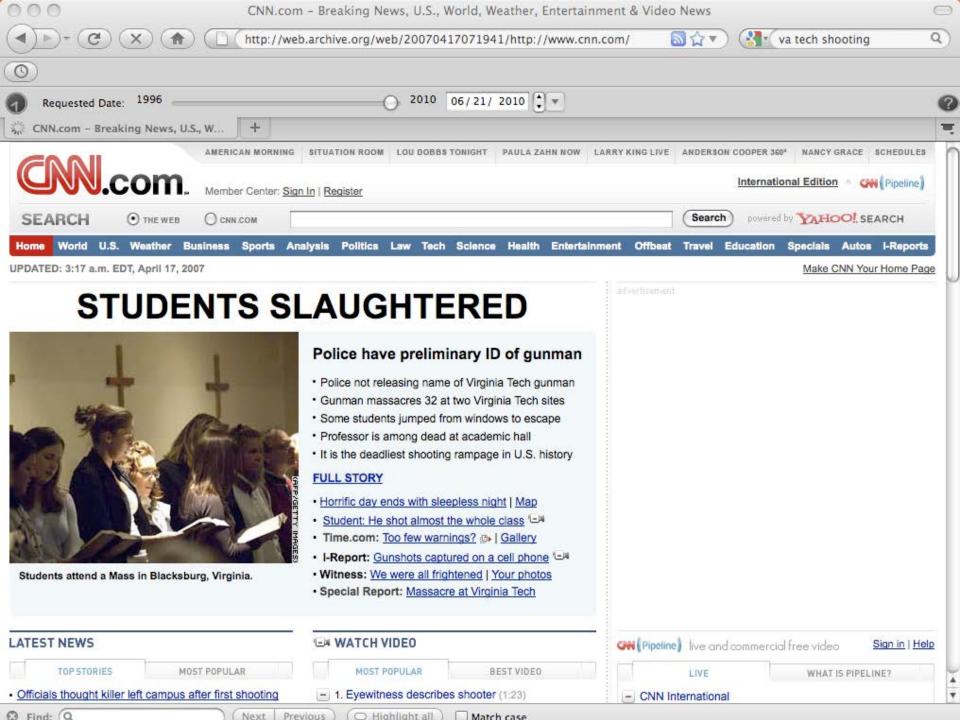




Archiving Moves At Hurricane Speed, Most News Stories Move Faster











Most of the Story, at Least as Conveyed by cnn.com, is Missing...

in this case, you can reconstruct the events with http://en.wikipedia.org/wiki/Virginia_Tech_massacre_timeline





How Much of The Web Is Archived?





Public Archives, ca. Late 2010 / Early 2011

Three categories of archives

- Internet Archive
- Search engine
- Other archives













The National Archives

UK







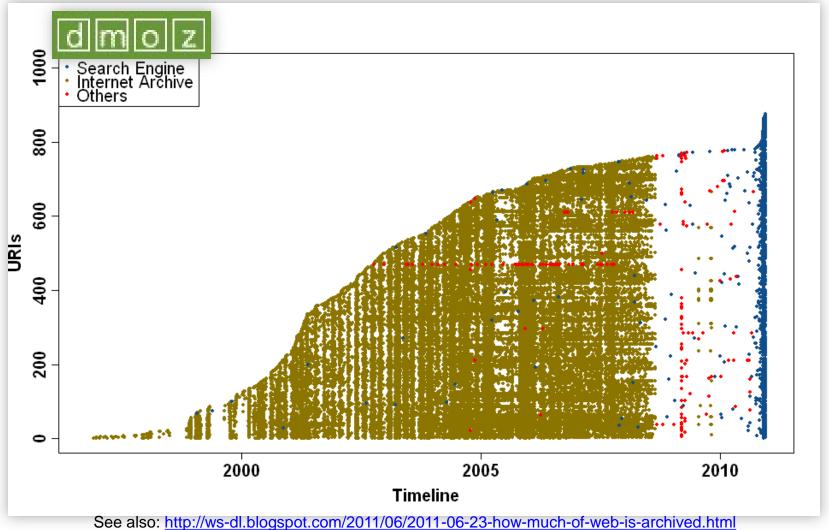


See also: http://arxiv.org/abs/1212.6177





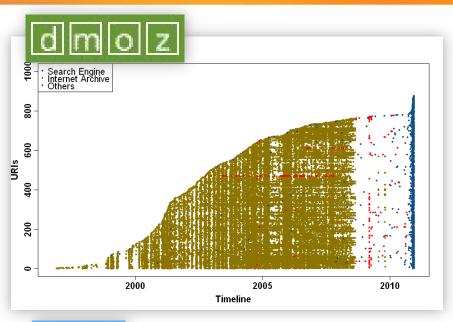
1000 URIs Ordered by First Observation Date

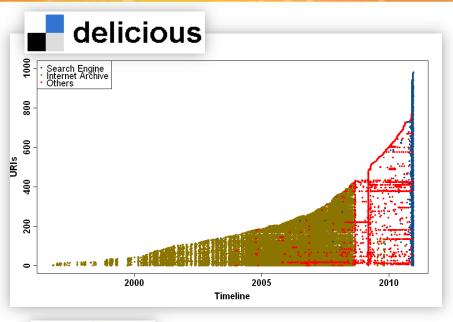


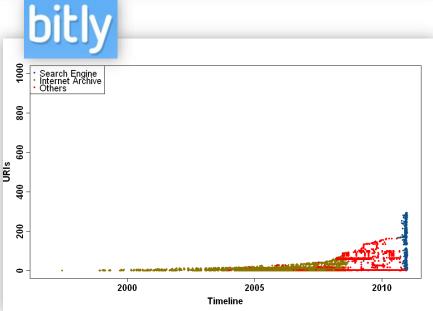
Wolfram DATA SUMMIT 2013

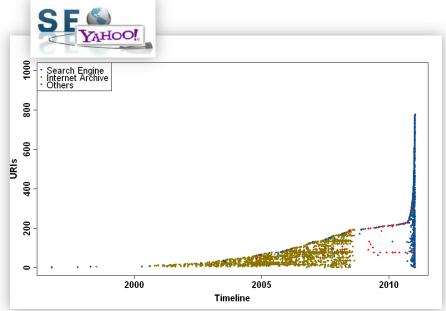










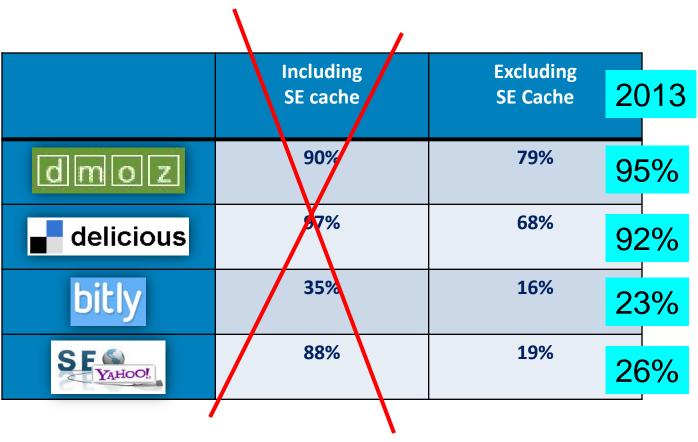


see also: http://ws-dl.blogspot.com/2013/04/2013-04-19-carbon-dating-web.html





How Much of the Web is Archived? It Depends on Which Web...



Changes since 2011: no more free SE APIs; greatly reduced IA quarantine period; 15 public web archives





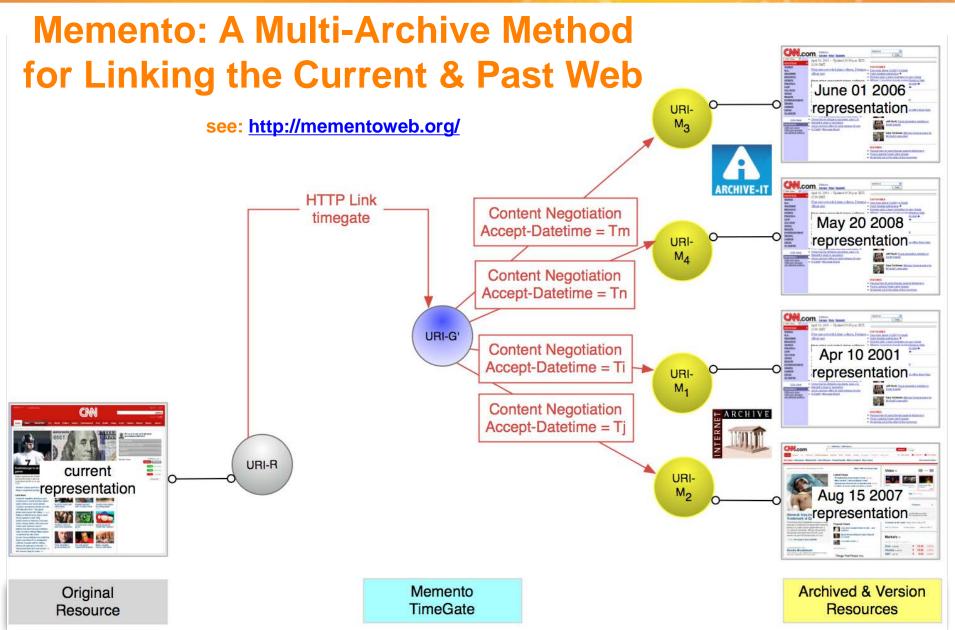
Long Tail of Archives



see also: http://www.cs.odu.edu/~mln/pubs/tpdl-2013/paper_134.pdf











So It's Been Archived, What Can Go Wrong?





Temporal Drift



August 27, 2005 11:16 a.m. EDT <u>link</u>





Temporal Drift: Now 3 Hours in the Past



August 27, 2005 11:16 a.m. EDT <u>link</u> August 27, 2005 8:00 a.m. EDT link





Temporal Drift: Now 17 Days in the Future



August 27, 2005 11:16 a.m. EDT <u>link</u>

8:00 a.m. EDT <u>link</u>

August 27, 2005 September 13, 2005 8:12 a.m. EDT <u>link</u>



Temporal Drift: Now 23 (or 6) Days in the Future



August 27, 2005 11:16 a.m. EDT link

8:00 a.m. EDT link

8:12 a.m. EDT link

August 27, 2005 September 13, 2005 September 19, 2005 8:25 a.m. EDT link

10+ clicks in the archive results in median drift of ~45 days (standard UI) or ~15 days with Memento. ~2% of the sessions have drift of > 1 year. see: http://www.cs.odu.edu/~mln/pubs/jcdl-2013/jcdl93-ainsworth.pdf

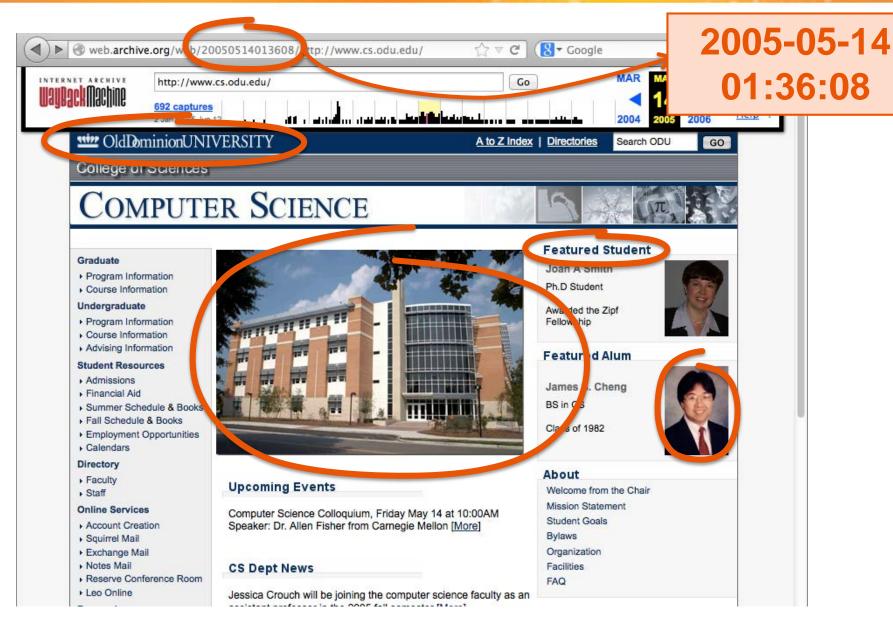




We Call the Drift in a Single Page "Temporal Spread"

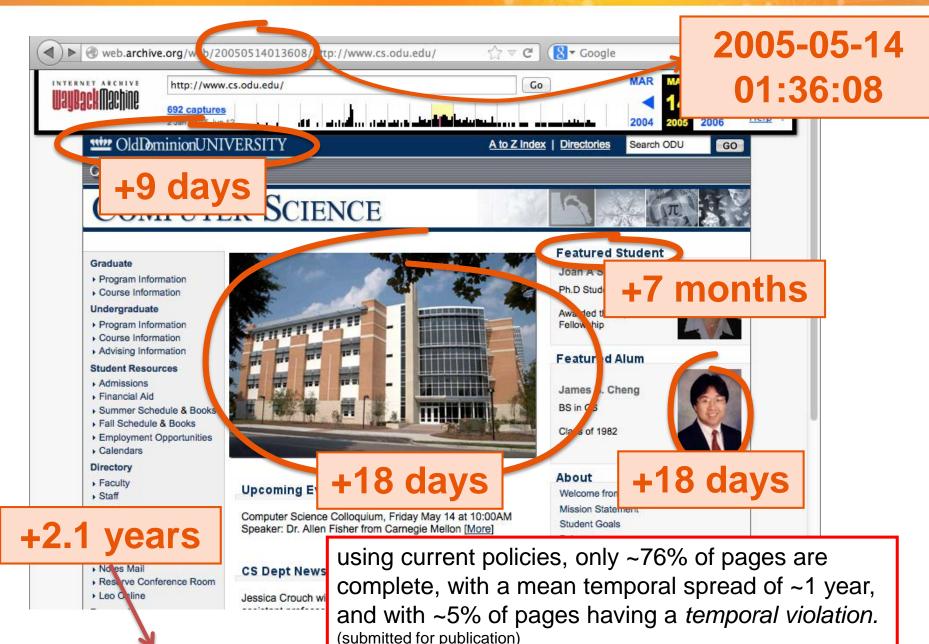
















Sometimes the Live Web "Leaks" Into the Archive...







see: http://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html



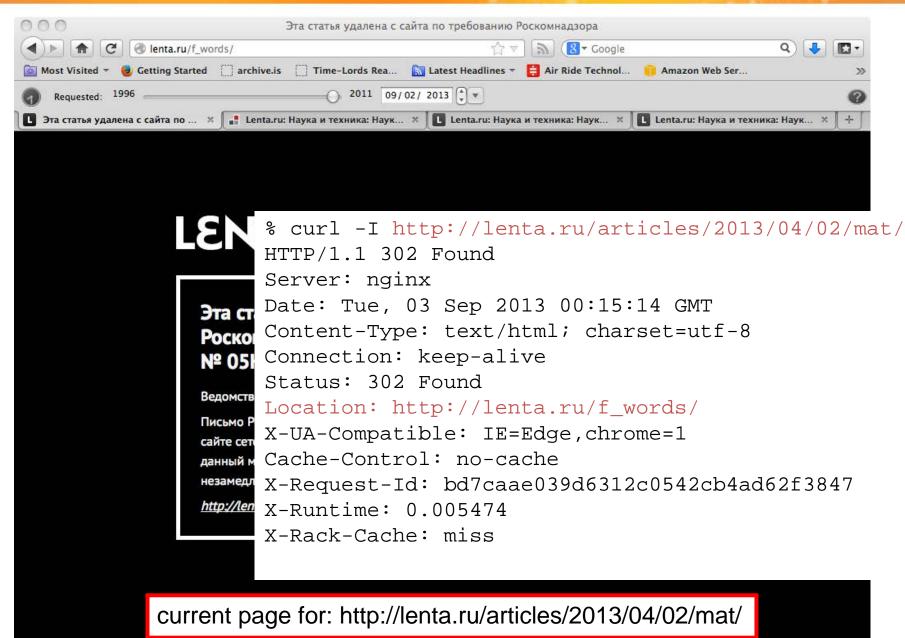


Quis Archiviet Ipsos Archives?

(thanks to webmaster@archive.is for this example)



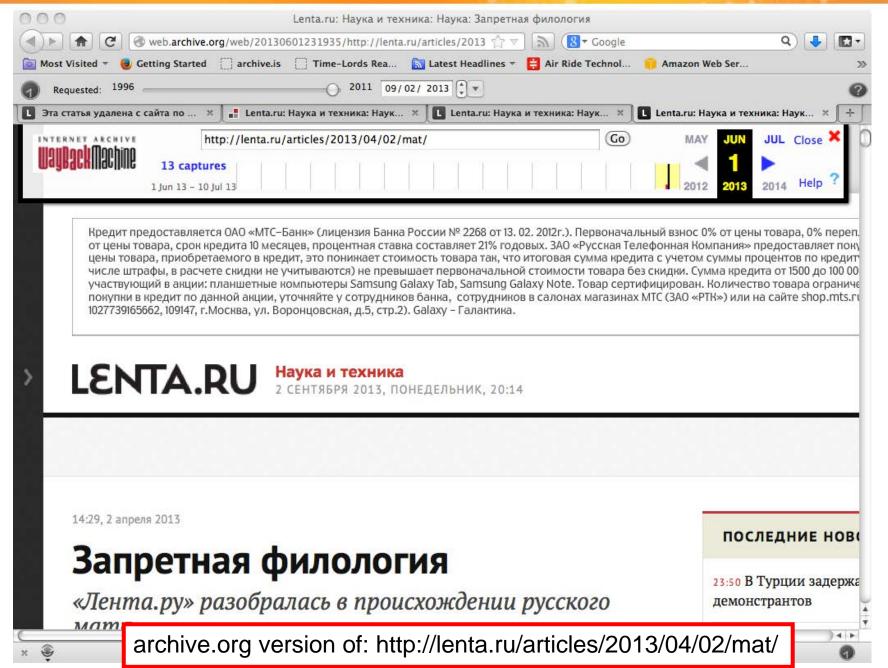






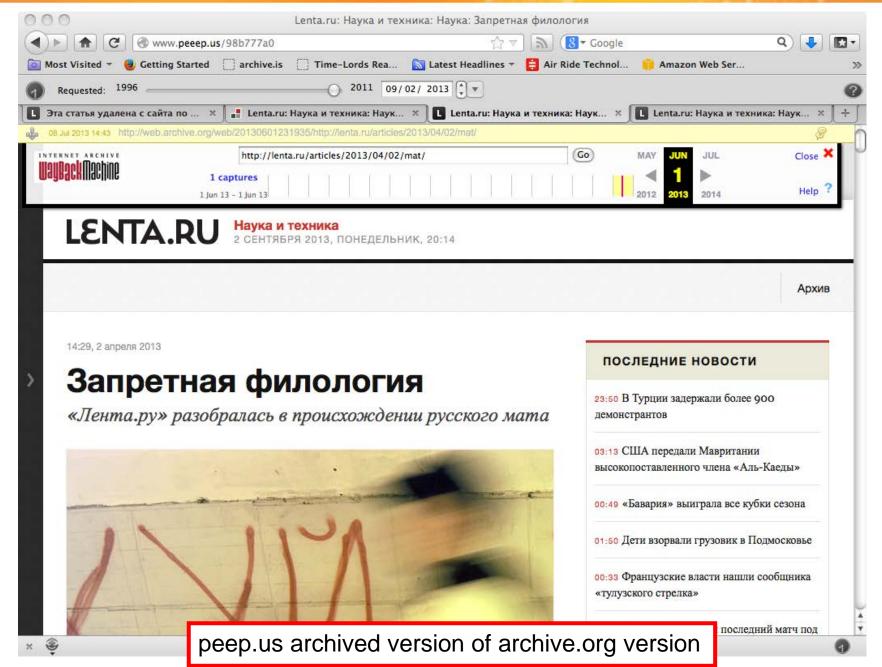






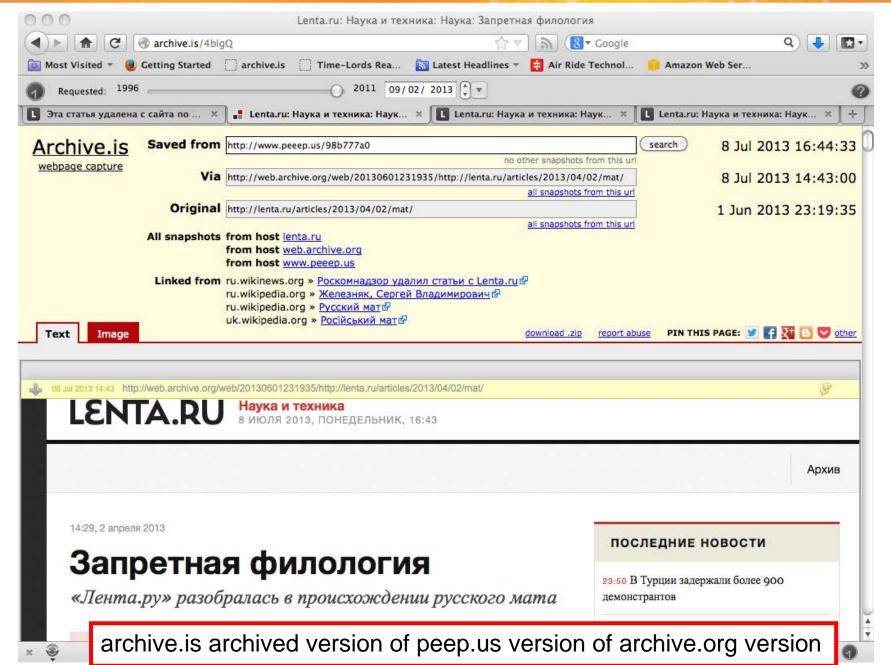














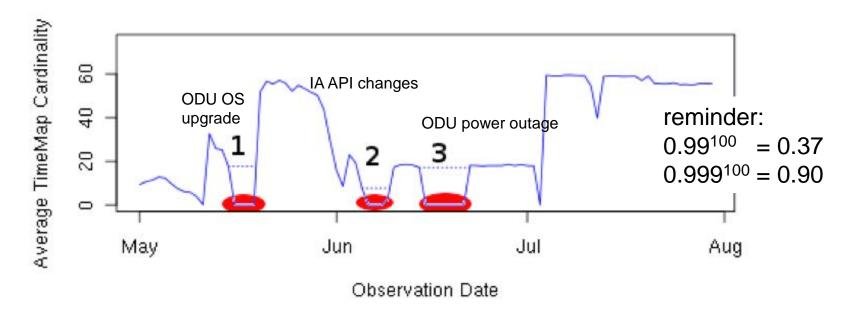


Why Make Lots of Copies?





Archives Are Subject to the Same Vagaries of Other Web Sites...



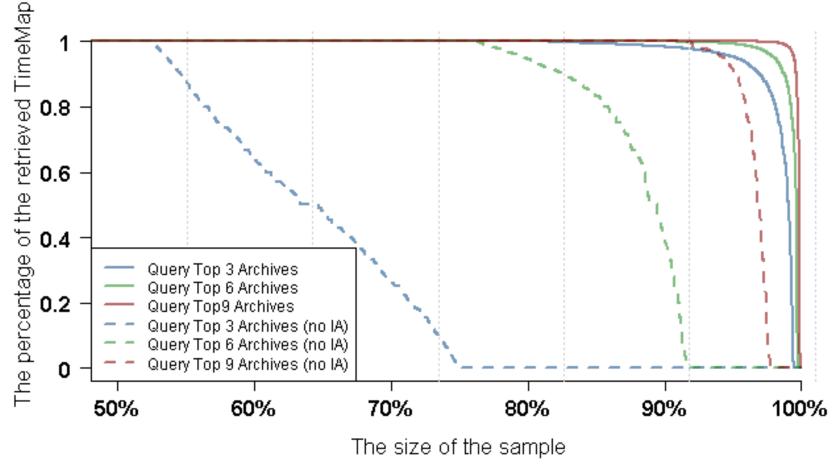
In a perfect world, this graph should be monotonically increasing. Memento allows simultaneous access to more archives, but this also means that at any given time, some archive(s) will be down.

see: http://arxiv.org/abs/1307.5685





Query Routing: Using Only Top-k Archives for URI Lookup Yields Good Results



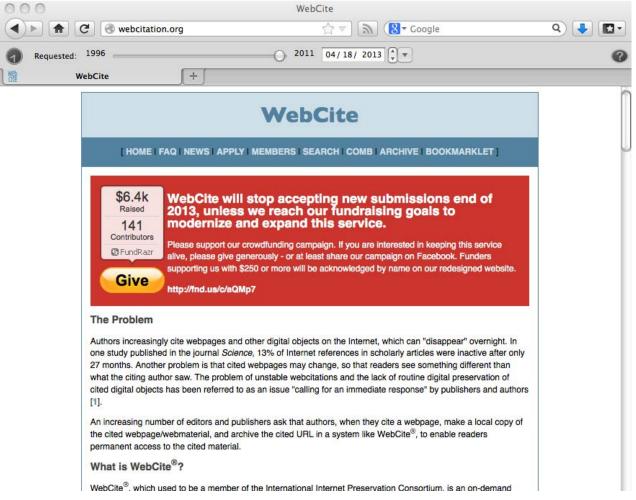
Even when there are 100s of archives, we only need to talk to a few.

see: http://www.cs.odu.edu/~mln/pubs/tpdl-2013/paper_134.pdf





What is the Economic Model for Archives?

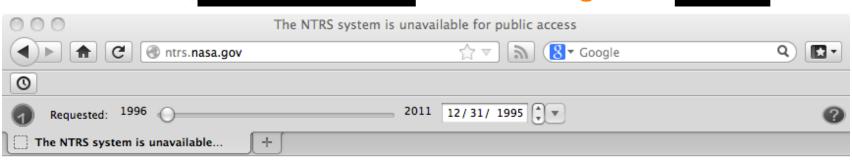


1TB endowment = ~\$4700: http://blog.dshr.org/2011/01/memento-marketplace-for-archiving.html





Houston, Here. The Eagle has



The NASA technical reports server will be unavailable for public access while the agency conducts a review of the site's content to ensure that it does not contain technical information that is subject to U.S. export control laws and regulations and that the appropriate reviews were performed. The site will return to service when the review is complete. We apologize for any inconvenience this may cause.









Summary

- We have a cultural mandate to preserve "obsolete data or resources"
 - however, we currently have limited discovery and replay tools
- We need lots of people making several copies of many things
 - Memento is the mechanism for accessing the long tail of archives