4-3-2016

Why We Need Multiple Archives

Michael L. Nelson

Old Dominion University, mnelson@odu.edu

Herbert Van de Sompel

Follow this and additional works at: http://digitalcommons.odu.edu/computerscience_presentations

Part of the Archival Science Commons

Recommended Citation


http://digitalcommons.odu.edu/computerscience_presentations/4

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.
Why We Need Multiple Archives

Michael L. Nelson & Herbert Van de Sompel
@phonedude_mln
@hvdseomp

Digital Preservation of Federal Information Summit
CNI Spring 2016
April 3, 2016
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
WHAT IF I TOLD YOU THERE ARE OTHER WEB ARCHIVES?
timetravel.mementoweb.org

e.g., bbc.co.uk in six different archives…
Seagal’s Law

A man with a watch knows what time it is. A man with two watches is never sure.

How to resolve conflicting archives?

Personalization, GeoIP, mobile vs. desktop, etc. means “the” page rarely exists, only “a” page.

Historicity of Web Archives

https://twitter.com/phonedude_mln/status/490171976389238784
Remember MH17?
https://twitter.com/phonedude_mln/status/490409622470393856

Michael L. Nelson @phonedude_mln · 22h
I’m often asked “why archive the (social) web?” B/c @internetarchive @archiveitorg have evidence of who downed #MH17 csmonitor.com/World/Europe/2...

Alex Sobakin @quicknquiet
@phonedude_mln @internetarchive @archiveitorg Yep. Perfect tool to produce "evidence" of any kind. Who would care about reliability?

1:15 AM - 19 Jul 2014

Don’t miss any updates from Alex Sobakin

Full name
Email
Password
Alex is now 404. Would multiple, independent archives have convinced him?

https://twitter.com/quicknquiet
A single archive is vulnerable

The Conservative Party has deleted speeches and press releases published on its website between 2000 and the 2010 general election.

The archive has also been hidden from search engines.

The story was spotted by Computer Weekly.

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.

Google says it acknowledges some people want 'right to be forgotten'

Corporation says it is trying to see things from 'a more European perspective' after receiving 220,000 requests for delisting in its search engine, but reveals 60% have been rejected
Economics working against archives

In the paper world in order to monetize their content the copyright owner had to maximize the number of copies of it. In the Web world, in order to monetize their content the copyright owner has to minimize the number of copies. Thus the fundamental economic motivation for Web content militates against its preservation in the ways that Herbert and I would like.

--David Rosenthal

http://blog.dshr.org/2015/02/the-evanescent-web.html
Who pays for those extra archives?

1TB endowment = ~$4700:  http://blog.dshr.org/2011/02/paying-for-long-term-storage.html
see also: http://blog.dshr.org/2011/01/memento-marketplace-for-archiving.html
Archives aren’t *magic* web sites
They’re just web sites.

If you used Mummify, you’re now left with a bunch of defunct, shortened links like:

https://mummify.it/XbmcMfE3
Don’t Throw Away the Original URL – Use Robust Links!

my robust link to the live web

my robust link to an archived version

Page Level Metadata Is The Least You Can Do

More examples / scenarios at: http://robustlinks.mementoweb.org/spec/
Archiving your internal stuff: Transactional Archiving

Never miss an update; archive your site as it is being viewed by users.

https://mementoweb.github.io/SiteStory/
Archiving your internal stuff: Heritrix & Wayback

Crawling your intranet: [http://www.dlib.org/dlib/january16/brunelle/01brunelle.html](http://www.dlib.org/dlib/january16/brunelle/01brunelle.html)
Crawling JS "stuff" will take 5X more storage: [http://arxiv.org/abs/1601.05142](http://arxiv.org/abs/1601.05142)

mementos of Mitre Intranet

“MiiTube” – Complete With Javascript leakage
JavaScript == the new deep web; use ResourceSync to make sure your URIs are exposed

<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
    xmlns:rs="http://www.openarchives.org/rs/terms/">
    <rs:ln rel="up"
        href="http://example.com/dataset1/capabilitylist.xml"/>
    <rs:md capability="resourcelist"
        at="2013-01-03T09:00:00Z"
        completed="2013-01-03T09:01:00Z"/>

    <url>
        <loc>http://example.com/res1</loc>
        <lastmod>2013-01-02T13:00:00Z</lastmod>
        <rs:md hash="md5:1584abdf8ebdc9802ac0c6a7402c03b6"
            length="8876"
            type="text/html"/>
    </url>

    <url>
        <loc>http://example.com/res2</loc>
        <lastmod>2013-01-02T14:00:00Z</lastmod>
        <rs:md hash="md5:1e0d5cb8ef6ba40c99b14c0237be735e
            sha-256:854f61290e2e197a11bc91063afce22e43f8ccc655237050ace766adc68dc784"
            length="14599"
            type="application/pdf"/>
    </url>
</urlset>

(AKA “Fancy SiteMaps”)
http://www.openarchives.org/rs/
When all else fails, justify project with: “web archiving is Big Data”