9-9-2010

My Point of View

Michael L. Nelson

Old Dominion University, mnelson@odu.edu

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

Part of the Archival Science Commons, and the Computer Sciences Commons

Recommended Citation

https://digitalcommons.odu.edu/computerscience_presentations/22

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.
My Point of View

Michael L. Nelson

Old Dominion University
Observations

• We are pretty good at archiving the web of five years ago, but not the web of today
• There are separate, shadow webs that we are not archiving
• Archiving should be a service with short-term utility
Ajax = #noarchive


My Point of View: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010
Reaching Out From the Archive


% grep Host: cnn-ia-headers | wc -l
288
% grep Host: cnn-ia-headers | grep -v archive.org | wc -l
117
% grep Host: cnn-ia-headers | grep -v archive.org | sort -u
Host: ad.doubleclick.net
Host: ads.adsonar.com
Host: ads.cnn.com
Host: aranet.vo.llnwd.net
Host: b.scorecardresearch.com
Host: bs.serving-sys.com
Host: cnn.dyn.cnn.com
Host: ds.serving-sys.com
Host: gdyn.cnn.com
Host: i.cdn.turner.com
Host: i2.cdn.turner.com
Host: js.adsonar.com
Host: metrics.cnn.com
Host: pix04.revsci.net
Host: s0.2mdn.net
Host: symbolcomplete.marketwatch.com
Host: www.adfusion.com
Reaching Through Time


My Point of View: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010
Embedded Resources

29 Mementos: http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://www.youtube.com/user/wichitarecordings
Personalized Resources

GET / HTTP/1.1
Host: bit.ly
User-Agent: Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.5; en-US; rv:1.9.2.10) Gecko/20100914 Firefox/3.6.10
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-us,en;q=0.5
Accept-Encoding: gzip, deflate
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Keep-Alive: 115
Connection: keep-alive
Cookie: _utma=126736798.4156477295523165000.1251253806.1285119293.1285122783.59; _bit=4c20df7a-003a5-07baf-91a08fa8; anon_u=cHN1X19jN2MwNjcxZC05MWNiLTQ3MmEtOGIxYy1hZDMyMWRlNzUyZTU=|1284997489|06ac0cefc8ac369e0f9849b5fdfb6e8d077d0c65; user=cGhvbmVkdWRl|1284997489|fdb7f02cabc3b44416f54d83f3237ec0f7bd9b5; __utmz=126736798.1280940647.33.1.utmcsr=(direct)|utmccn=(direct)|utmcmd=(none); _chartbeat2=ciuph6qrs06tn6w7; _xsrf=49bc661fc02845b3bcbe975d7c2f28de; __utmb=126736798.3.10.1285122783; __utmc=126736798
Geolocated Resources

% curl -I http://www.craigslist.org
HTTP/1.1 302 Found
Set-Cookie: cl_b=12851300231056905752;path=/;domain=craigslist.org;expires=01 Jan 2038 00:00:00 GMT
Location: http://geo.craigslist.org/

% curl -I http://geo.craigslist.org/
HTTP/1.1 302 Found
Content-Type: text/html; charset=iso-8859-1
Connection: close
Location: http://norfolk.craigslist.org
Date: Wed, 22 Sep 2010 04:33:56 GMT
Set-Cookie: cl_b=12851300363085180962;path=/;domain=craigslist.org;expires=01 Jan 2038 00:00:00 GMT
Server: Apache

% traceroute geo.craigslist.org
traceroute to geo.craigslist.org (208.82.236.208), 64 hops max, 40 byte packets
1  *  *  *
2  10.5.120.1 (10.5.120.1)  9.959 ms  23.004 ms  13.208 ms
3  nrfksysr02-atm151208.hr.hr.cox.net (68.10.8.117)  10.056 ms  10.561 ms  19.970 ms
4  nrfkdsrj01-ge500.0.rd.hr.cox.net (68.10.14.13)  11.142 ms  20.618 ms  10.293 ms
5  ashbbprj02-ae4.0.rd.as.cox.net (68.1.1.232)  15.368 ms  68.854 ms  20.153 ms
6  xe-3-0-0.cr2.dca2.above.net (64.125.26.241)  18.963 ms  23.674 ms  32.977 ms
7  xe-2-2-0.cr2.iah1.us.above.net (64.125.30.53)  46.201 ms  56.156 ms  46.783 ms
8  xe-1-1-0.mpr4.phx2.us.above.net (64.125.28.73)  82.616 ms  82.289 ms  84.383 ms
9  * 64.124.178.62.allocated.above.net (64.124.178.62)  80.893 ms  78.786 ms
10  511.ae9.ecore1p.craigslist.org (208.82.236.208)  80.968 ms  91.470 ms  80.110 ms
11  www.craigslist.org (208.82.236.208)  80.968 ms  91.470 ms  80.110 ms

My Point of View: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010
Social Resources

http://www.flickr.com/photos/mic_n_2_sugars/84882320/
0 Mementos: http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://farm1.static.flickr.com/37/84882320_67fc8915d5_z.jpg

My Point of View: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010
Shadow Web: Mobile

Where I've been the past week, when mostly offline: 800 miles of road, 20 miles on horseback, ancient Southwest ruins

@O'Reilly We've offered that, and people don't really want the combo.
About 12 hours ago via Sesamatic Desktop in reply to @timoreilly

@timoreilly My interview with @greenbiz on how the web is a sustainability platform http://gramee.toO6r0 (I'll be at their Innovation Forum in Oct) about 12 hours ago

@timoreilly Sanofi Aventis announces diabetes meter that connects to the #iPhone http://ibgstar.com http://bit.ly/dnsZQZ #sensors via @josephafaber

0 Mementos: http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://mobile.twitter.com/timoreilly
# Shadow Web: Mobile


Shadow Web: Linked Data

http://en.wikipedia.org/wiki/DJ_Shadow

http://dbpedia.org/resource/DJ_Shadow

Accept: application/rdf+xml

Accept: text/html

0 Mementos: http://mementoproncy.cs.odu.edu/aggr/timemap/link/http://dbpedia.org/data/DJ_Shadow
0 Mementos: http://mementoproncy.cs.odu.edu/aggr/timemap/link/http://dbpedia.org/page/DJ_Shadow

My Point of View: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010
Archive Discovery

% curl -I http://dbpedia.org/resource/DJ_Shadow
HTTP/1.1 303 See Other
Date: Wed, 22 Sep 2010 04:13:16 GMT
Content-Type: text/html; charset=UTF-8
Connection: keep-alive
Server: Virtuoso/06.02.3128 (Linux) x86_64-generic-linux-glibc25-64 VDB
Accept-Ranges: bytes
Location: http://dbpedia.org/page/DJ_Shadow
Content-Length: 0
Set-Cookie: uid=wm2BOkyZglwm1zEBBv2+Ag==; expires=Sat, 02-Oct-10 04:13:16 GMT; domain=dbpedia.org; path=/
P3P: policyref="w3c/p3p.xml", CP="CUR ADM OUR NOR STA NID"

% curl -I http://dbpedia.org/page/DJ_Shadow
HTTP/1.1 200 OK
Date: Wed, 22 Sep 2010 04:23:15 GMT
Content-Type: text/html; charset=UTF-8
Connection: keep-alive
Vary: Accept-Encoding
Server: Virtuoso/06.02.3128 (Linux) x86_64-generic-linux-glibc25-64 VDB
Expires: Wed, 29 Sep 2010 03:39:43 GMT
Content-Length: 60711
Set-Cookie: uid=wm2BOkyZhLOGQzDjB4JHAg==; expires=Sat, 02-Oct-10 04:23:15 GMT; domain=dbpedia.org; path=/
P3P: policyref="w3c/p3p.xml", CP="CUR ADM OUR NOR STA NID"
Accept-Ranges: bytes

DBpedia archive now hosted @ LANL:

My Point of View: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010
Decontextualized Resources
from these we can create time-based:

- indexes
- IDF values
- PageRank

Tagging

http://www.delicious.com/url/4c858ce7188a51bfb3b80b3011cbed8
http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://www.delicious.com/url/4c858ce7188a51bfb3b80b3011cbed8
Tweeting

http://twitter.com/#search?q=%23thecribs
http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://twitter.com/#search?q=%23thecribs (false Mementos!)
Searching

http://www.google.com/trends?q=the+cribs
Analytics

Non-Archiving Made Easy…

#NoLOC.org

Keep Your Tweets From Being Archived Forever

Every single Twitter tweet will be archived forever by the US government. Use #NoLOC to set your tweets free!

Starting 10/10/2010, Twitter will begin transferring all of your tweets to the Library of Congress for archiving. Tweets that are 24 weeks old (6 months) or older will be saved forever, and you will no longer be able to delete them. By using #NoLOC.org, you continue to use Twitter like normal, but when you tweet turns 23 weeks old, we’ll delete it automatically.

Just sign in with Twitter to use NoLOC.org, and all of your tweets that include the hashtag #noloc will not be sent to the Library of Congress for archiving. It’s that simple.

Did you know that every single tweet will be available forever in the Library of Congress? #noloc
Batch Recovery For Sites

In order to recover a lost website, please fill out the information below:

First name: Dwight
Last name: Schrupe
Email address: dschrupe@dundermiflin.com
Website URL to recover: http://battlestar-galactica.com/

Recover entire website? Yes  No, only recover the single page

Repositories to use:
- Internet Archive
- Google
- Yahoo!
- Live Search

Use Windows filenames? Yes  No

Submit Clear

* A required field.
** Only the latest resources from the Internet Archive will be returned unless dates are specified.

http://warrick.cs.odu.edu/

My Point of View: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010
Real-Time Recovery for URIs

Synchronicity - www.cs.odu.edu/~mklein/

My Point of View: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010
RT, @, #, bit.ly, tiny.cc… users are willing to endure an appalling level of syntax if there is a clear and present benefit…
Closing Thoughts

Preservation not for privileged priesthood
http://doi.acm.org/10.1145/1592761.1592794
http://booktwo.org/notebook/wikipedia-historiography/

Don't dessicate resources; leave them on the web

no more hoary stories about format obsolescence:

Archiving as branded service, not infrastructure
http://blog.dshr.org/2010/06/icdl-2010-keynote.html

Endless metadata is not preservation…