9-9-2010

My Point of View

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

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

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

Recommended Citation
http://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
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


My Point of View: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010
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=cH1X19jN2MwNjcxZC05MWNiLTQ3MmEtOGIxYY1hZDMyMWRINzc1OTU=[1284997489]06ac0cefc8ac369e0f9849b5fdffbe8d077d0c65; user=cGhvbmVkdWRl[1284997489]f02c44416f45d83f3237ec0f7bd9b5; _umtza=126736798.1280940647.33.1.utmcsr=(direct)utmccn=(direct)utmcmd=(none); _chartbeat2=ciup6qrs6tn6w7; _xsrf=49bc661fc02845b3bcbe975d7c2f28de; __utmb=126736798.3.10.1285122783; __utmc=126736798
Geolocated Resources

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/
http://farm1.static.flickr.com/37/84882320_67fc8915d5_z.jpg (Last-Modified: 10 Jan 2006…)
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

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

<table>
<thead>
<tr>
<th>Year</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages</td>
<td>0</td>
<td>30</td>
<td>2734</td>
<td>231</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Archived Results from Jan 01, 1996 - latest


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

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

http://dbpedia.org/resource/DJ_Shadow

Accept: text/html

Accept: application/rdf+xml

http://dbpedia.org/page/DJ_Shadow

http://dbpedia.org/data/DJ_Shadow


0 Mementos: http://mementoproyxy.cs.odu.edu/aggr/timemap/link/http://dbpedia.org/data/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/1.1 200 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
Content-Length: 60711
Set-Cookie: uid=wm2B0KyZglwm1zEBBv2+Ag==; 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"

DBpedia archive now hosted @ LANL:
http://mementoarchive.lanl.gov/dbpedia/timegate
http://dbpedia.org/page/DJ_Shadow
http://mementoarchive.lanl.gov/dbpedia/timemap/rdf
http://dbpedia.org/page/DJ_Shadow

% 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
Server: Virtuoso/06.02.3128 (Linux) x86_64-generic-linux-glibc25-64 VDB
Expires: Wed, 29 Sep 2010 03:39:43 GMT
Link: <http://dbpedia.org/data/DJ_Shadow.rdf>; rel="alternate"; type="application/rdf+xml"; title="Structured Descriptor Document (RDF/XML format)",
<http://dbpedia.org/data/DJ_Shadow.n3>; rel="alternate"; type="text/n3"; title="Structured Descriptor Document (N3/Turtle format)",
<http://dbpedia.org/data/DJ_Shadow.json>; rel="alternate"; type="application/json"; title="Structured Descriptor Document (RDF/JSON format)",
<http://dbpedia.org/data/DJ_Shadow.atom>; rel="alternate"; type="application/atom+xml"; title="OData (Atom+Feed format)",
<http://dbpedia.org/resource/DJ_Shadow>; rev="describedby",
<http://mementoarchive.lanl.gov/dbpedia/timegate/http://dbpedia.org/page/DJ_Shadow>; rel="timegate"
Content-Length: 0
Set-Cookie: uid=wm2B0KyZhL0gLqZqDjB4JHBa==; 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
Decontextualized Resources…
from these we can create time-based:
- indexes
- IDF values
- PageRank

Tagging

http://www.delicious.com/url/4c858ce7188a51bf3b80b3011cbed8
http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://www.delicious.com/url/4c858ce7188a51bf3b80b3011cbed8
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 #nocloc 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:* Schrute
Email address:* dschrute@dundermifflin.com
Website URL to recover:* http://battlestar-galactica.com/

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

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

Use Windows filenames? * Yes   No   Help

* 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/](http://www.cs.odu.edu/~mklein/)

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

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

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

no more hoary stories about format obsolescence:

Endless metadata is not preservation…