Investigating Reference Rot in Web-Based Scholarly Communication

Herbert Van de Sompel  
Los Alamos National Laboratory  
@hvdsomp

Martin Klein  
Los Alamos National Laboratory  
@mart1nkle1n

http://hiberlink.org #hiberlink  
http://mementoweb.org #memento

Hiberlink is funded by the Andrew W. Mellon Foundation
Hiberlink Project Partners

• **Los Alamos National Laboratory:**
  • **Research Library:** Martin Klein, Robert Sanderson, Herbert Van de Sompel

• **University of Edinburgh:**
  • **Edina:** Peter Burnhill, Neil Mayo, Muriel Mewissen, Christine Rees, Tim Stickland, Richard Wincewicz
  • **Language Technology Group:** Beatrice Alex, Claire Grover, Richard Tobin, Ke “Adam” Zhou

• **Funding:** Andrew W. Mellon Foundation
Acknowledgments

• Primary datasets: arXiv, Chesapeake Project, Elsevier, PubMed Central, PLoS, … (many more to come)

• Secondary datasets: Ex Libris, MS Academic, SerialsSolutions

• Technology support: CrossRef Labs, CrossRef Prospect, Elsevier

• Liaisons: archive.is, CrossRef, Internet Archive, Old Dominion University Web Science & Digital Library Research Group, perma.cc
Reference Rot
Problem Domain

- Web-based scholarly communication links to, references, Web resources:
  - Formal citing of scholarly resources
  - Referencing “Web at Large” resources needed or created in research activities e.g. project websites, software, ontologies, workflows, online debate, slides, blogs, videos, etc.
Problem Domain

- Links to web resources are subject to Reference Rot:
  - Link Rot: Link stops working, e.g. HTTP 404
  - Content Decay: Linked content changes over time
## References in Web-Based Scholarly Communication

<table>
<thead>
<tr>
<th></th>
<th>To Scholarly Resources</th>
<th>To Web at Large Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Link Rot</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Content Decay</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

an increasingly blurry boundary
## References in Web-Based Scholarly Communication

<table>
<thead>
<tr>
<th></th>
<th>To Scholarly Resources</th>
<th>To Web at Large Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Link Rot</strong></td>
<td>DOI, HTTP version of DOI</td>
<td></td>
</tr>
<tr>
<td><strong>Content Decay</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### References in Web-Based Scholarly Communication

<table>
<thead>
<tr>
<th></th>
<th>To Scholarly Resources</th>
<th>To Web at Large Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Link Rot</strong></td>
<td>DOI, HTTP version of DOI</td>
<td></td>
</tr>
<tr>
<td><strong>Content Decay</strong></td>
<td>Fixity of content</td>
<td></td>
</tr>
</tbody>
</table>
## References in Web-Based Scholarly Communication

<table>
<thead>
<tr>
<th></th>
<th>To Scholarly Resources</th>
<th>To Web at Large Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Link Rot</strong></td>
<td>DOI, HTTP version of DOI</td>
<td></td>
</tr>
<tr>
<td><strong>Content Decay</strong></td>
<td>Fixity of content</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Archiving: CLoCKSS, LoCKSS, Portico, Keepers Registry, …</td>
<td></td>
</tr>
</tbody>
</table>
### References in Web-Based Scholarly Communication

<table>
<thead>
<tr>
<th>Link Rot</th>
<th>To Scholarly Resources</th>
<th>To Web at Large Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOI, HTTP version of DOI</td>
<td>Archiving: CLoCKSS, LoCKSS, Portico, Keepers Registry, …</td>
<td></td>
</tr>
</tbody>
</table>

There are issues here too, see David Rosenthal blog post http://blog.dshr.org/2013/11/patio-perspectives-at-anadp-ii.html
# References in Web-Based Scholarly Communication

## Table: Link Rot and Content Decay

<table>
<thead>
<tr>
<th></th>
<th>To Scholarly Resources</th>
<th>To Web at Large Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Link Rot</strong></td>
<td>DOI, HTTP version of DOI</td>
<td></td>
</tr>
<tr>
<td><strong>Content Decay</strong></td>
<td>Fixity of content</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Archiving: CLoCKSS, LoCKSS, Portico, Keepers Registry, …</td>
<td></td>
</tr>
</tbody>
</table>

Hiberlink focus
References to “Web at Large” Resources

• Hiberlink focuses on the wide variety of web resources needed or created in research activities

• These resources:
  • Are not necessarily under the custodianship of a party that cares about long term integrity, access
  • Do not necessarily have the same sense of fixity that e.g. journal articles have

• Reference Rot makes it impossible to adequately recreate the temporal context for scholarly discourse
Rethinking Scholarly Communication

Building the System that Scholars Deserve

Herbert Van de Sompel
Los Alamos National Laboratory, Research Library
<herbertv@lanl.gov>

Sandy Payette
Cornell University, Computing and Information Science
<payette@cs.cornell.edu>

John Erickson
Hewlett-Packard Laboratories, Digital Media Systems Lab
<john.erickson@hp.com>

Carl Lagoze
Cornell University, Computing and Information Science
<lagoze@cs.cornell.edu>

Simeon Warner
Cornell University, Computing and Information Science
<simeon@cs.cornell.edu>
References


Hiberlink: Investigating Reference Rot

- Hiberlink explores references to Web at Large resources:
  - Quantifies Reference Rot
  - Explores potential solutions to Reference Rot
  - Focuses on links in electronic journal articles
    - But has the big picture in mind: dynamic, interdependent, web-based scholarly assets
  - See Herbert Van de Sompel, From the Version of Record to a Version of the Record, CNI Spring 2013 plenary talk - http://www.youtube.com/watch?v=fhrGS-QbNVA
# References in Web-Based Scholarly Communication

<table>
<thead>
<tr>
<th>Link Rot</th>
<th>To Scholarly Resources</th>
<th>To Web at Large Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Decay</td>
<td>DOI, HTTP version of DOI</td>
<td></td>
</tr>
<tr>
<td>Fixity of content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Archiving: CLoCKSS, LoCKSS, Portico, Keepers Registry, …</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Is it worth our time to study this?
Articles Increasingly Link to Web Resources

URIs extracted from PubMed papers – links to Web at Large resources
SIDEBAR
In Supreme Court Opinions, Web Links to Nowhere
By ADAM LIPTAK
Published: September 23, 2013

WASHINGTON — Supreme Court opinions have come down with a bad case of link rot. According to a new study, 49 percent of the hyperlinks in Supreme Court decisions no longer work.

This can sometimes be amusing. A link in a 2011 Supreme Court opinion about violent video games by Justice Samuel A. Alito Jr. now leads to a mischievous error message.

“Aren’t you glad you didn’t cite to this Web page?” it asks. “If you had, like Justice Alito did, the original content would have long since disappeared and someone else might have come along and purchased the domain in order to make a comment about the transience of linked information in the Internet age.”

The prankster has a point. The modern Supreme Court opinion is increasingly built on sand.

Hyperlinks are a huge and welcome convenience, of course, said Jonathan Zittrain, who teaches law and computer science at Harvard and who prepared the study with Kendra Albert, a law student there. “Things are readily accessible,” he said, “until they aren’t.”

Quantifying Reference Rot
Quantifying Reference Rot

- Reference Rot has been studied before:
  - For the web at large
  - For scholarly communication
  - For government documents

- What is different with Hiberlink?
  - Investigates Reference Rot not just link rot, i.e. includes the aspect of changing content not just rotting links
  - Investigates coverage of referenced resources in web archives
  - Operates at a massive scale regarding number of journal articles, referenced URIs, web archive lookups
<table>
<thead>
<tr>
<th>STUDY Author (Date)</th>
<th>Year of Publication of Citations</th>
<th># URIs</th>
<th>#URIs looked up in web archives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lawrence (2001)</td>
<td>1993-1999</td>
<td>67,577</td>
<td>-</td>
</tr>
<tr>
<td>Davis (2002)</td>
<td>1999-2001</td>
<td>688</td>
<td>-</td>
</tr>
<tr>
<td>Wren (2004)</td>
<td>1994-2002</td>
<td>1,630</td>
<td>-</td>
</tr>
<tr>
<td>Sellitto (2005)</td>
<td>1995-2003</td>
<td>1,043</td>
<td>-</td>
</tr>
<tr>
<td>Goh (2005)</td>
<td>1997-2003</td>
<td>2,516</td>
<td>-</td>
</tr>
<tr>
<td>Dimitrova (2007)</td>
<td>2000-2003</td>
<td>1,126</td>
<td>-</td>
</tr>
<tr>
<td>McCown (2005)</td>
<td>1995-2004</td>
<td>4,387</td>
<td>-</td>
</tr>
<tr>
<td>Parker (2007)</td>
<td>2002-2005</td>
<td>1,229</td>
<td>-</td>
</tr>
<tr>
<td>Duda (2008)</td>
<td>1997-2005</td>
<td>2,100</td>
<td>-</td>
</tr>
<tr>
<td>Falagas (2007)</td>
<td>2003-2006</td>
<td>1,417</td>
<td>-</td>
</tr>
<tr>
<td>Russell (2008)</td>
<td>1999-2006</td>
<td>510</td>
<td>-</td>
</tr>
<tr>
<td>Wren (2008)</td>
<td>1994-2007</td>
<td>6,154</td>
<td>-</td>
</tr>
<tr>
<td>Moghaddam (2010)</td>
<td>1995-2008</td>
<td>1,761</td>
<td>1,761</td>
</tr>
<tr>
<td><strong>Sanderson (2011)</strong></td>
<td><strong>1993-2010</strong></td>
<td><strong>162,052</strong></td>
<td><strong>162,052</strong></td>
</tr>
</tbody>
</table>

Quantifying Reference Rot - Methodology

article → HTTP URI extraction → Referenced URIs + article metadata

not article → Filtered URIs + article metadata

URI-R → Web Archive lookup → Memento TimeMap

Live Web lookup → HTTP transactions

Memento TimeMap → HTTP transactions
HTTP URI extraction

Referenced URIs + article metadata

not article

Filtered URIs + article metadata

URI-R

Web Archive lookup

Live Web lookup

Memento TimeMap

HTTP transactions

Memento TimeMap

HTTP transactions

Memento TimeMap

HTTP transactions

HTTP transactions

HTTP transactions

HTTP transactions
- HTTP HEAD on referenced URI-R
- Follow redirects up to a maximum of 50
- Record HTTP transaction chain
- If HTTP transaction chain ends with 2XX status code: **Exists**
- If HTTP transaction chain does not end with 2XX: **!Exist**
• Lookup in web archives via a Memento Aggregator that covers among others Internet Archive, Archive-It, archive.is, British Library web archive, UK National Archives web archive, Icelandic web archive
• Obtain TimeMap per URI
  • If TimeMap does not exist: Archived
  • If TimeMap exists, select Memento URI-M closest to article publication date
  • HTTP HEAD on URI-M
  • Follow archived redirects up to a maximum of 50
  • Record HTTP transaction chain
  • If HTTP transaction chain ends 2XX: Archived
  • If HTTP transaction chain does not end with 2XX: Archived
Data used for analysis

HTTP URI extraction

Referenced URIs + article metadata

not article

Filtered URIs + article metadata

URI-R

Web Archive lookup

Live Web lookup

Memento TimeMap

HTTP transactions

Memento TimeMap

HTTP transactions

Memento TimeMap

HTTP transactions
Quantifying Reference Rot – Early Results

[Graph showing the percentage of references that exist (31.2%), are archived (16.8%), archived within 30 days (40.7%), within 14 days (11.3%) and within 1 day.]

[Graph showing the amount of citations over time.]
Study: PubMed Central Corpus 01/1997 - 12/2012

• Articles processed: 494,785
• Articles that contain Web at Large URIs: 176,527
• References to Web at Large URIs: 557,432
• Unique referenced Web at Large URIs: 327,782
Percentage Exists & Archived Referenced URIs

- Exists & Archived: 31.2%
- !Exists & Archived: 16.8%
- Exists & !Archived: 11.3%
- !Exists & !Archived: 40.7%

URIs extracted from PubMed papers – links to Web at Large resources
Percentage Exists & Archived in 30 Day Window

- Exists & Archived: 23%
- !Exists & Archived: 16.7%
- Exists & !Archived: 55.2%
- !Exists & !Archived: 5.1%

URIs extracted from PubMed papers – links to Web at Large resources
Percentage Exists & Archived in 15 Day Window

- Exists & Archived: 24.6%
- Exists & Archived: 12.4%
- Exists & !Archived: 59.5%
- !Exists & !Archived: 3.5%

URIs extracted from PubMed papers – links to Web at Large resources
Percentage Exists & Archived in 07 Day Window

- Exists & Archived: 25.8%
- !Exists & Archived: 8.8%
- Exists & !Archived: 2.3%
- !Exists & !Archived: 63.1%

URIs extracted from PubMed papers – links to Web at Large resources
Percentage Exists & Archived in 01 Day Window

- **27.9%** Exists & Archived
- **0.9%** !Exists & Archived
- **0.2%** Exists & !Archived
- **71%** !Exists & !Archived

URIs extracted from PubMed papers – links to Web at Large resources
Percentage of !Exists per Year

URIs extracted from PubMed papers – links to Web at Large resources
Percentage of !Exists, Archived per Year

URIs extracted from PubMed papers – links to Web at Large resources
Solving Reference Rot
## References in Web-Based Scholarly Communication

<table>
<thead>
<tr>
<th></th>
<th>To Scholarly Resources</th>
<th>To Web at Large Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Link Rot</strong></td>
<td>DOI, HTTP version of DOI</td>
<td></td>
</tr>
<tr>
<td><strong>Content Decay</strong></td>
<td>Fixity of content</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Archiving: CLoCKSS, LoCKSS, Portico, Keepers Registry, …</td>
<td></td>
</tr>
</tbody>
</table>
Addressing Content Decay

- Aim for a more pro-active approach to collect snapshots of web resources (likely to be) referenced in scholarly communication.

- A system that hosts resources that are likely to be referenced in scholarly communication can create snapshots of itself by:
  - Using CMS, wikis, datawikis with solid versioning mechanisms
  - Subscribing to on-demand self web archiving service
  - Using transactional web archives, cf. SiteStory

- Referenced resources can be web archived on-demand:
  - By authors during note taking, authoring
  - By platforms involved in the publication process, e.g. archiving linked resources at the time of manuscript submission
## References in Web-Based Scholarly Communication

<table>
<thead>
<tr>
<th></th>
<th>To Scholarly Resources</th>
<th>To Web at Large Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link Rot</td>
<td>DOI, HTTP version of DOI</td>
<td></td>
</tr>
<tr>
<td>Content Decay</td>
<td>Fixity of content</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Archiving: CLoCKSS, LoCKSS, Portico, Keepers Registry, …</td>
<td>Web archiving Content Versioning Systems Self archiving</td>
</tr>
</tbody>
</table>

Herbert Van de Sompel, Martin Klein – Hiberlink
CNI Fall 2013, Washington, DC, December 9 2013
In Supreme Court Opinions, Web Links to Nowhere

By ADAM LIPTAK
Published: September 23, 2013

WASHINGTON — Supreme Court opinions have come down with a bad case of link rot. According to a new study, 49 percent of the hyperlinks in Supreme Court decisions no longer work.

This can sometimes be amusing. A link in a 2011 Supreme Court opinion about violent video games by Justice Samuel A. Alito Jr. now leads to a mischievous error message.

“Aren’t you glad you didn’t cite to this

Click link to blog post
http://blogs.law.harvard.edu/futureoftheinternet/2013/09/22/perma/
Perma: Scoping and addressing the problem of “link rot”

September 22nd, 2013  by z  Published in Future of the Internet  23 Comments

Kendra Albert, Larry Lessig and I are finishing up a study of link rot, available at http://papers.ssrn.com/abstract=2329161. Link rot is the phenomenon by which material we link to on the distributed Web vanishes or changes beyond recognition over time. (Wiki discusses link rot [here.]) This is a particular problem for academic scholarship, which is increasingly linking out to the Web rather than more formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up...
Search and find Mementos in Internet Archive for
http://blogs.law.harvard.edu/futureoftheinternet/2013/09/22/perma/
<table>
<thead>
<tr>
<th>Oldest shot</th>
<th>Newest shot</th>
<th>List of URLs, ordered from newer to older</th>
</tr>
</thead>
</table>

Search and find a Memento in archive.is for
date each site was last visited, and its clerk keeps a hard copy of those materials. In an interview, Ms. Liebler said the court should do more.

“It’s a half measure to put a piece of paper in a court file,” she said. “This is the Supreme Court, and it’s their responsibility to make these things available.”

The United States Court of Appeals for the Ninth Circuit, in San Francisco, could serve as a model. It maintains an electronic archive of what it calls “webcites” in the PDF format.

Professor Zittrain and his colleagues are at work on a more ambitious solution, Perma.cc, a platform built and run by a consortium of law libraries. It allows writers and editors to capture and fix transient information on the Web with a new, permanent link.

The project is initially focused on legal scholarship. And there is no reason, Professor Zittrain said, why it could not also work for the Supreme Court.

A version of this article appears in print on September 24, 2013, on page A13 of the New York edition with the headline: In Supreme Court Opinions, Web Links to Nowhere.

Click perma.cc link to Memento of blog post http://perma.cc/0Hg62eLdZ3T
Perma: Scoping and addressing the problem of “link rot”

Perma.cc

September 22nd, 2013  |  by z  |  Published in Future of the Internet  |  23 Comments

Kendra Albert, Larry Lessig and I are finishing up a study of link rot, available at http://kendra.batistin.com/archives/2826161. Link rot is the phenomenon by which material we link to on the distributed Web vanishes or changes beyond recognition over time. (Wiki discusses link rot here.) This is a particular problem for academic scholarship, which is increasingly linking out to the Web rather than more formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up with sources that can’t be checked or followed up upon by readers.

We found that half of the links in all Supreme Court opinions no longer work. And more than 70% of the links in such journals as the Harvard Law Review (in that case measured from 1999 to 2012), currently don’t work. As time passes, the number of non-working links increases.

Our work builds on other great link rot studies such as that by Raizel Liebler and June Liebert in the Yale Journal of Law and Technology, available here (PDF).

In response, the Harvard Library Innovation Lab has pioneered a project to unite libraries so that link rot can be mitigated. We are joined by about thirty law libraries around the world to start Perma.cc, which will allow those libraries on direction of authors and journal editors to store permanent caches of otherwise ephemeral links. Libraries are the ideal partners for this task: they think on a long timescale, they take user trust and service seriously; and they are non-commercial. You can see more about the system at perma.cc. The amazing Internet Archive has lent its archiving engine to the effort, and Instapaper has generously provided an alternative with its parsed Web pages to be accessed. Both efforts have kindly ensured that the link system of

Receive Memento from perma.cc
http://perma.cc/0Hg62eLDZ3T

Herbert Van de Sompel, Martin Klein – Hiberlink
CNI Fall 2013, Washington, DC, December 9 2013
Hrm.

Wayback Machine doesn't have that page archived.
Herbert Van de Sompel, Martin Klein – Hiberlink
CNI Fall 2013, Washington, DC, December 9 2013

Search and do not find Mementos in archive.is for
http://perma.cc/0Hg62eLdZ3T
What Happened?

• Good news: The number of archived copies of the blog post was increased by pro-actively creating a Memento in perma.cc

• Bad news: The possibility of finding Mementos for the blog post in other web archives was undermined by replacing the Original URI-R with the Memento URI-M
  • The Memento URI-M is a key in only one archive
  • The Original URI-R is a key in all web archives

• Using the Memento URI-M in a link requires the permanent existence/uptime of the archive that issued it

• One link rot problem was replaced by another …
Web Archives Less Permanent than Permanent?

WebCite

$11.6k Raised
267 Contributors

Give

WebCite will stop accepting new submissions end of 2013, unless we reach our fundraising goals to modernize and expand this service.

Please support our crowdfunding campaign. If you are interested in keeping this service alive, please give generously - or at least share our campaign on Facebook. Funders supporting us with $250 or more will be acknowledged by name on our redesigned website.

http://fnd.us/c/aQMpz

The Problem

Authors increasingly cite webpages and other digital objects on the Internet, which can "disappear" overnight. In one study published in the journal Science, 13% of Internet references in scholarly articles were inactive after only 27 months. Another problem is that cited webpages may change, so that readers

http://webcitation.org
Web Science and Digital Libraries Research Group
Research and Teaching Updates from the Web Science and Digital Libraries Research Group at Old Dominion University.

Thursday, November 21, 2013
2013-11-21: The Conservative Party Speeches and Why We Need Multiple Web Archives

Michael L. Nelson
@phonedude_mln

@Conservatives put speeches in Streisand's house: computerweekly.com/blogs/public-s...
@UKWebArchive: webarchive.org.uk/ukwa/target/10... via @lljohnston @hhockx

11:43 AM - 13 Nov 2013
3 RETWEETS

Web Science and Digital Libraries
This blog is used to communicate research and education updates from the Web Science and Digital Libraries Research Group at Old Dominion University.

Please see the web page of Michael L. Nelson for more information:
http://www.cs.odu.edu/~mln/

Contributors
Haroun SalahFideen
Hawood Alam
Web Archives Less Permanent than Permanent?

What To Do?

- Need an approach for referencing archived resources that supports lookups in many web archives, not just one

- Since the Original URI-R is a key in all web archives, the linking approach needs to necessarily include it

- Hence, two URIs are required:
  - The Original URI-R
  - The Memento URI-M, e.g. the perma.cc URI

- But a link in HTML only carries one URI!
  - It is understandable that the Memento URI-M is used for the link: the approach works with existing web infrastructure
  - Yet, an approach to address link rot that itself is subject to link rot is … err… problematic
The Missing Link Proposal

- Extend the link to the Original URI-R with temporal context:
  - Memento URI-M in a specific archive
  - Dates:
    - date of page that contains the link
    - date of the link, cf. “accessed at” in citations of web resources

- Provide the Original URI-R and the temporal context in a machine-actionable manner so it can be used by user and machine agents to retrieve Mementos from various web archives

http://mementoweb.org/missing-link/
The Missing Link Proposal

```html
<META content="....." itemprop="datePublished">

<a href="....." versiondate="....." versionurl="....."></a>
```

to prior version of resource with unknown version URI

to current version of resource

to prior version of resource with known version URI

http://mementoweb.org/missing-link/
How to Make Missing Link Happen?

• The existing approach works out of the box but is problematic

• Missing Link requires infrastructure changes but generally contributes to increased web persistence:
  • HTML
    • META for page date: no problem, already in use
    • Attributes for <a> to convey URI-M and link date:
      • data– extensibility mechanism in HTML5 can be used but is not intended for cross-site applications
      • In 1995, HTML had the URN attribute for <a> as a means to address web persistence concerns
  • Browser, tool support
# References in Web-Based Scholarly Communication

<table>
<thead>
<tr>
<th>Link Rot</th>
<th>To Scholarly Resources</th>
<th>To Web at Large Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOI, HTTP version of DOI</td>
<td>Missing Link proposal</td>
<td></td>
</tr>
<tr>
<td>Content Decay</td>
<td>Fixity of content</td>
<td>-</td>
</tr>
<tr>
<td>Archiving: CLoCKSS, LoCKSS, Portico, Keepers Registry, …</td>
<td>Web archiving Content Versioning Systems Self archiving</td>
<td></td>
</tr>
</tbody>
</table>
Demo: Application Using Temporal Context for Links
Investigating Reference Rot in Web-Based Scholarly Communication

Herbert Van de Sompel  
Los Alamos National Laboratory  
@hvdsomp

Martin Klein  
Los Alamos National Laboratory  
@mart1nkle1n

http://hiberlink.org #hiberlink  
http://mementoweb.org #memento

Hiberlink is funded by the Andrew W. Mellon Foundation
Demo: Application Using Temporal Context for Links

In Supreme Court Opinions, Web Links to Nowhere

WASHINGTON -- Supreme Court opinions have come down with a bad case of link rot. According to a new study, 49 percent of the hyperlinks in Supreme Court decisions no longer work.

A link in a 2011 Supreme Court opinion about violent video games by Justice Samuel A. Alito Jr. now leads to a miscellaneous error message.

"Aren't you glad you didn't cite to this," Bob Zaleski, the Web editor at The New York Times, said Tuesday. Wikipedia has the article."

Perma: Scoping and addressing the problem of "link rot"

September 22nd, 2013 1 by 2 Published in Future of the Internet | 23 Comments

Kendra Albert, Larry Lessig and I are finishing up a study of link rot, available at http://papers.sem.com/abstract=0529981. Link rot is the phenomenon by which material we link to on the distributed Web vanishes or changes beyond recognition over time. (Wikipedia discuses link rot here.) This is a particular problem for academic scholarship, which is increasingly linking out to the Web rather than more formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up with sources that can't be checked or followed up upon by readers.

We found that half of the links in all Supreme Court opinions no longer work. And more than 70% of the links in such journals as the Harvard Law Review (in that case measured from 1999 to 2012) currently don't work. As time passes, the number of non-working links increase.
Application Using Temporal Context for Links

• Memento for Chrome is an application that uses Original URI-R and dates to access Mementos in various web archives
  • Memento around the date selected in user interface calendar
  • Most recently archived Memento
Memento Time Travel for Chrome

Install the Memento extension for Chrome

Memento Time Travel for Chrome

http://www.youtube.com/watch?v=0_70lQPOOIg
http://www.youtube.com/watch?v=WtZHKeFwjzk
Application Using Temporal Context for Links

- An experimental version of Memento for Chrome also uses Missing Link information (Original URI-R, URI-M, and dates) to access Mementos in various web archives:
  - Memento around the date selected in user interface calendar
  - Most recently archived Memento
  - Memento around the date of the page that contains the link
  - Memento around the date of the link
  - Memento URI-M in a specific archive

- A Memento client is just one example of an application that can use temporal context provided for links. Other applications, including search engines, can use it too
In Supreme Court Opinions, Web Links to Nowhere

WASHINGTON — Supreme Court opinions have come down with a bad case of link rot. According to a new study, 49 percent of the hyperlinks in Supreme Court decisions no longer work.

This can sometimes be amusing. A link in a 2011 Supreme Court opinion about violent video games by Justice Samuel A. Alito Jr. now leads to a mischievous error message.

"Aren’t you glad you didn’t cite to this Web page?" it asks. "If you had, like,

Link in NYT was:
<a href="http://blogs.law.harvard.edu/futureoftheinternet/2013/09/22/perma/">
Changed to:
<a href="http://blogs.law.harvard.edu/futureoftheinternet/2013/09/22/perma/"
data-versionurl="http://perma.cc/0Hg62eLdZ3T"
In Supreme Court Opinions, Web Links to Nowhere

By ADAM LIPTAK
Published: September 23, 2013

WASHINGTON — Supreme Court opinions have come down with a bad case of link rot. According to a new study, hyperlinks in Supreme Court decisions no longer function.

This can sound like a problem. But in a 2009 article, Samuel A. Alito Jr., a mischievous justice, wrote: “Aren’t you the one who isn’t supposed to click the link?”

enabler: <a href="URI-R">
Perma: Scoping and addressing the problem of “link rot”

September 22nd, 2013 by z Published in Future of the Internet 23 Comments

Kendra Albert, Larry Lessig and I are finishing up a study of link rot, available at http://papers.ssrn.com/abstract=2329161. Link rot is the phenomenon by which material we link to on the distributed Web vanishes or changes beyond recognition over time. (Wiki discusses link rot here.) This is a particular problem for academic scholarship, which is increasingly linking out to the Web rather than more formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up with sources that can’t be checked or followed up upon by readers.

Receive Memento from archive.is, Nov 24 2013
In Supreme Court Opinions, Web Links to Nowhere

By ADAM LIPTAK
Published: September 23, 2013

WASHINGTON — Supreme Court opinions have come down with a bad case of link rot. According to a new study about violent Samuel A. Alito, Jr., mischievous "Aren't you going to try to make the Web page?" it asks. "If you had like

Right Click Link Get at page date
http://blogs.law.harvard.edu/futureoftheinternet/2013/09/22/perma/enabler: <a href="URI-R"> & <META itemprop="datePublished" content="2013-09-23">
Perma: Scoping and addressing the problem of “link rot”

September 22nd, 2013 | by z | Published in Future of the Internet | 3 Comments

Kendra Albert, Larry Lessig and I are finishing up a study of link rot, available at http://papers.ssrn.com/abstract=2329161. Link rot is the phenomenon by which material we link to on the distributed Web vanishes or changes beyond recognition over time. (Wiki discusses link rot here.) This is a particular problem for academic scholarship, which is increasingly linking out to the Web rather than more formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up with sources that can’t be checked or followed up upon by readers.

We found that half of the links in all Supreme Court opinions no longer work. And more than 70% of the links in such journals as the Harvard Law Review (in that case measured from 1999 to 2012), currently don’t work. As time passes, the number of non-working links increases.

Our work builds on other great link rot studies such as that by Raizel Liebler and June Liebert in the Yale Journal of Law and Technology, available here (PDF).

In response, the Harvard Library Innovation Lab has pioneered a project to unite libraries so that link rot can be mitigated. We are joined by about thirty law libraries around the world to start Perma.cc, which will allow those...
In Supreme Court Opinions, Web Links to Nowhere

By ADAM LIPTAK
Published: September 23, 2013

WASHINGTON — Supreme Court opinions have come down with a bad case of link rot. According to a new study, many hyperlinks in Supreme Court decisions no longer work.

This can sometimes be because of the judge’s references—such as a link in a 2013 opinion by Justice Samuel A. Alito Jr. to a paper that didn’t exist anymore. But it is also sometimes because the court’s Web page is missing.

“ Aren’t you going to have a Web page?” it asks. “ If you had like a new study?”

Right Click Link Get from perma.cc
http://blogs.law.harvard.edu/futureoftheinternet/2013/09/22/perma/

enabler: <a href=“URI-R” data-versionurl=“URI-M”>
Receive Memento from perma.cc, Oct 2 2013
http://perma.cc/0Hg62eLdZ3T
Link in NYT was:

<a href="http://perma.cc/0Hg62eLdZ3T">

Changed to:

<a href="http://blogs.law.harvard.edu/futureoftheinternet/2013/09/22/perma/">

data-versionurl="http://perma.cc/0Hg62eLdZ3T"
date each site was last visited, and its clerk keeps a hard copy of those materials. In an interview, Ms. Liebler said the court should do more.

“It’s a half measure to put a piece of paper in a court file,” she said. “This is the Supreme Court, and it’s their responsibility to make these things available.”

The United States Court of Appeals for the Ninth Circuit, in San Francisco, could serve as a model. It maintains an electronic archive of what it calls “websites” in the PDF format.

Professor Zittrain and his colleagues are at work on a more ambitious solution, Perma.cc, a platform built and run by a consortium of law libraries. It allows writers and editors to capture and fix transient information on the Web with a new, permanent link.

The project is initially focused on legal scholarship. And there is no reason, Professor Zittrain said, why it could not also work for the Supreme Court.

A version of this article appears in print on September 24, 2013, on page A13 of the New York edition with the article "The Court and Its Websites." In Supreme Court Opinions, Web Links to Nowhere.


Get Free E-mail Alerts on These Topics

Computers and the Internet

Supreme Court

 Longer-Term Deal With Iran Faces Major Challenges

Get Version Overview
Perma: Scoping and addressing the problem of “link rot”

September 22nd, 2013

by z | Published in Future of the Internet | 23 Comments

Kendra Albert, Larry Lessig and I are finishing up a study of link rot, available at http://papers.ssrn.com/abstract=2329161. Link rot is the phenomenon by which material we link to on the distributed Web vanishes or changes beyond recognition over time. (Wiki discusses link rot here.) This is a particular problem for academic scholarship, which is increasingly linking out to the Web rather than more formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up with sources that can’t be checked or followed up upon by readers.

We found that half of the links in all Supreme Court opinions no longer work. And more than 70% of the links in such journals as the Harvard Law Review (in that case measured from 1999 to 2012), currently don’t work. As time passes, the number of non-working links increases.
Perma: Scoping and addressing the problem of “link rot”

September 22nd, 2013 | by z | Published in Future of the Internet | 23 Comments

Kendra Albert, Larry Lessig and I are finishing up a study of link rot, available at http://papers.ssrn.com/abstract=2329161. Link rot is the phenomenon by which material we link to on the distributed Web vanishes or changes beyond recognition over time. This is a particular problem for academic scholarship, which is increasingly linking to the Web rather than more formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up with sources that can’t be checked or followed up upon by readers.

We found that half of the links in all Supreme Court opinions no longer work. And more than 70% of the links in such journals as the Harvard Law Review (in that case measured from 1999 to 2012), currently don’t work. As time passes, the number of non-working links increases.

Click Link (done on November 25 2013)
http://en.wikipedia.org/wiki/Link_rot

enabler: <a href="URI-R">
Link rot

From Wikipedia, the free encyclopedia


Link rot (or linkrot), also known as link death or link breaking, is an informal term for the process by which hyperlinks (either on individual websites or the Internet in general) point to web pages, servers or other resources that have become permanently unavailable. The phrase also describes the effects of failing to update out-of-date web pages that clutter search engine results. A link that does not work any more is called a broken link, dead link or dangling link.

Contents [hide]
1 Causes
2 Prevalence
3 Discovering
4 Combating
   4.1 Server side
   4.2 User side
   4.3 Web archiving
   4.4 Authors citing URLs

Receive Page
http://en.wikipedia.org/wiki/Link_rot
External links [edit]

- Future-Proofing Your URLs
- Warrick - a tool for recovering lost websites from the Internet Archive and search engine caches
- UndeadLinks.com - user-contributed databases of moved URLs
- W3C Link Checker
- mod_brokenlink - Apache module that reports broken links.
- Perma.cc - Perma.cc archives a copy of the referenced content, and generates a link to an unalterable hosted instance of the site.

Categories: Uniform resource locator | Data quality

This page was last modified on 22 October 2013 at 13:27.
Perma: Scoping and addressing the problem of “link rot”

September 22nd, 2013 | by z | Published in Future of the Internet | 23 Comments

Kendra Albert, Larry Lessig and I are finishing up a study of link rot, available at http://papers.ssrn.com/abstract=2329161. Link rot is the phenomenon by which material we link to on the distributed Web vanishes or changes beyond recognition over time. (Wiki discusses link rot here.)

A particular problem for academic scholarship, which is increasingly linking out to the Web rather than to formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up with sources that can’t be checked or followed up upon by readers.

We found that half of the links in all Supreme Court opinions no longer work. And more than 70% of links in such journals as the Harvard Law Review (in that case measured from 1999 to 2012), could not be retrieved. As time passes, the number of non-working links increases.

Right Click Link Get at page date
http://en.wikipedia.org/Link_rot
enabler: <a href="URI-R"> & <META itemprop="datePublished" content="2013-09-22">
Link rot

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by Rchandra (talk | contribs) at 03:38, 3 September 2013. It may differ significantly from the current revision.

(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)


Link rot (or linkrot), also known as link death or link breaking, is an informal term for the process by which hyperlinks (either on individual websites or the Internet in general) point to web pages, servers or other resources that have become permanently unavailable. The phrase also describes the effects of failing to update out-of-date web pages that clutter search engine results. A link that does not work any more is called a broken link, dead link or dangling link.

Contents [hide]
1 Causes
2 Prevalence
3 Discovering
4 Combating

Receive Page
External links

- Future-Proofing Your URLs
- Warrick - a tool for recovering lost websites from the Internet Archive and search engine caches
- UndeadLinks.com - user-contributed databases of moved URLs
- W3C Link Checker
- mod_brokenlink - Apache module that reports broken links.

Categories: Uniform resource locator | Data quality

This version of the page has been revised.
Besides normal editing, the reason for revision may have been that this version contains factual inaccuracies, vandalism, or material not compatible with the Creative Commons Attribution-ShareAlike License.

Privacy policy  About Wikipedia  Disclaimers  Contact Wikipedia  Developers  Mobile view

Scroll down in page
Does not show Perma.cc link, added October 22 2013, a month after the blog post
formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up with sources that can't be checked or followed up upon by readers.

We found that half of the links in all Supreme Court opinions no longer work. And more than 70% of the links in such journals as the Harvard Law Review (in that case measured from 1999 to 2012), currently don't work. As time passes, the number of non-working links increases.

Our work builds on other great link rot studies such as that by Raizel Liebler and June Liebert in the Yale Journal of Law and Technology, available here (PDF).

In response, the Harvard Library Innovation Lab has pioneered a project to unite libraries so that link rot can be mitigated. We are joined by about thirty law libraries around the world to start Perma.cc, which will allow those libraries on direction of authors and journal editors to store permanent caches of otherwise ephemeral links. Libraries are the ideal partners for this task: they think on a long timescale; they take user trust and service seriously; and they are non-commercial. You can see more about the system at perma.cc. The amazing Internet Archive has lent its archiving engine to the effort, and Instapaper has generously provided an alternative path to parse Web pages to be saved. CloudFlare has kindly ensured that the the system at Perma.cc can scale with use.

We’re grateful to these many institutions and people who have come together to help make the Web work for the ages — the only way this can work is as a peer effort.

Perma’s founding partners are:

* Ponce Law Library, Washington College of Law, American University

Link in blog was:

<a href="http://librarylab.law.harvard.edu">Link in blog was:</a>

Changed (for fun) to:

<a href="http://librarylab.law.harvard.edu" data-versiondate="2010-09-22">Changed (for fun) to:</a>
formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up with sources that can’t be checked or followed up upon by readers.

We found that half of the links in all Supreme Court opinions no longer work. And more than 70% of the links in such journals as the Harvard Law Review (in that case measured from 1999 to 2012), currently don’t work. As time passes, the number of non-working links increases.

Our work builds on other great link rot studies such as that by Raizel Liebler and June Liebert in the Yale Journal of Law and Technology, available here (PDF).

In response, the Harvard Library Innovation Lab has pioneered a project to unite libraries so that link rot can be mitigated. We are joined by about thirty law libraries around the world to start Perma.cc, which will allow those libraries on direction of authors and journal editors to store permanent caches of otherwise ephemeral links. Libraries are the ideal partners for this task: they think on a long timescale; they take user trust and service seriously; and they are non-commercial. You can see more about the system at perma.cc. The amazing Internet Archive has lent its archiving engine to the effort, and Instapaper has generously provided an alternative path to parse Web pages to be saved. CloudFlare has kindly ensured that the the system at Perma.cc can scale with use.

We’re grateful to these many institutions and people who have come together to help make the Web work for the ages — the only way this can work is as a peer effort.

Perma’s founding partners are:

- Pence Law Library, Washington College of Law, American University

Click Link (done on November 25 2013)
http://librarylab.law.harvard.edu
enabler: <a href="URI-R">
formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up with sources that can't be checked or followed up upon by readers.

We found that half of the links in all Supreme Court opinions no longer work. And more than 70% of the links in such journals as the Harvard Law Review (in that case measured from 1999 to 2012), currently don't work. As time passes, the number of non-working links increases.

Our work builds on other great link rot studies such as that by Raizel Liebler and June Liebert in the Yale Journal of Law and Technology, available here (PDF).

In response, the Harvard Library Innovation Lab can be mitigated. We are joined by about thirty allow those libraries on direction of authors and ephemeral links. Libraries are the ideal partner trust and service seriously; and they are non-commercial partners. The amazing Internet Archive has lent its archive provided an alternative path to parse Web pages system at Perma.cc can scale with use.

We’re grateful to these many institutions and people for the ages — the only way this can work is as a peer effort.

Perma’s founding partners are:

- Pence Law Library, Washington College of Law, American University

Right Click Link Get at page date
http://librarylab.law.harvard.edu

enabler: <a href="URI-R"> & <META itemprop="datePublished" content="2013-09-22">
formal, library-curated sources. That kind of linking makes clear sense, but having materials easily accessible right until they vanish means that academic work (government documents, such as judicial opinions) can end up with sources that can't be checked or followed up upon by readers.

We found that half of the links in all Supreme Court opinions no longer work. And more than 70% of the links in such journals as the Harvard Law Review (in that case measured from 1999 to 2012), currently don't work. As time passes, the number of non-working links increases.

Our work builds on other great link rot studies such as that by Raizel Liebler and June Liebert in the Yale Journal of Law and Technology, available here (PDF).

In response, the Harvard Library Innovation Lab can be mitigated. We are joined by about thirty allow those libraries on direction of authors and ephemeral links. Libraries are the ideal partners of link rot. The amazing Internet Archive has lent its archive, provided an alternative path to parse Web pages, and system at Perma.cc can scale with use.

We're grateful to these many institutions and people for the ages — the only way this can work is as a peer effort.

Perma's founding partners are:

* Pence Law Library, Washington College of Law, American University

Right Click Link Get at link date

http://librarylab.law.harvard.edu

enabler: <a href="URI-R" data-versiondate="2010-09-22">
The Harvard Library Innovation Laboratory at the Harvard Law School

*Hello. We're a new lab exploring the future of libraries - because, well, we think their future needs exploring.*

**Projects**

---

Receive [Memento from Internet Archive, Sep 18 2010](http://web.archive.org/web/20100918025331/http://librarylab.law.harvard.edu)
Bottom Line: A Link Leads to Many Times and Archives

<table>
<thead>
<tr>
<th>current</th>
<th>&lt;a href=&quot;...&quot;&gt;&lt;/a&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>REPRESENTATION</td>
<td>ENABLER</td>
</tr>
<tr>
<td>most recently archived</td>
<td>&lt;a href=&quot;...&quot;&gt;&lt;/a&gt;</td>
</tr>
<tr>
<td>selected calendar date</td>
<td>user interface element</td>
</tr>
<tr>
<td>date on page</td>
<td>&lt;meta itemprop=&quot;datePublished&quot; content=&quot;...&quot;&gt;</td>
</tr>
<tr>
<td>date on link</td>
<td>&lt;a href=&quot;...&quot; versiondate=&quot;...&quot;&gt;&lt;/a&gt;</td>
</tr>
<tr>
<td>archived URI on link</td>
<td>&lt;a href=&quot;...&quot; versionurl=&quot;...&quot;&gt;&lt;/a&gt;</td>
</tr>
</tbody>
</table>

http://an.example.org

http://mementoweb.org/missing-link/