

# Reference Rot and Link Decoration

**Martin Klein**

UCLA

[martinklein0815@gmail.com](mailto:martinklein0815@gmail.com)

[@mart1nkle1n](#)



# Hiberlink Team

- Los Alamos National Laboratory
  - Research Library: (Martin Klein), (Robert Sanderson), Harihar Shankar, **Herbert Van de Sompel**
- University of Edinburgh
  - Edina: **Peter Burnhill**, Neil Mayo, Muriel Mewissen, Christine Rees, Tim Strickland, Richard Wincewicz
  - Language Technology Group: Beatrix Alex, **Claire Grover**, Colin Matheson, Richard Tobin, (Ke “Adam” Zhou)
- Funding: Andrew W. Mellon Foundation



OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

# Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot

Martin Klein , Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, Richard Tobin

Published: December 26, 2014 • DOI: 10.1371/journal.pone.0115253

13 Saves	0 Citations
11,560 Views	307 Shares

Article	Authors	Metrics	Comments	Related Content
---------	---------	---------	----------	-----------------

Download PDF

Print Share

Abstract

- Introduction
- Methods
- Results
- Discussion
- Supporting Information
- Acknowledgments
- Author Contributions
- References

- Reader Comments (0)
- Media Coverage (1)
- Figures

## Abstract

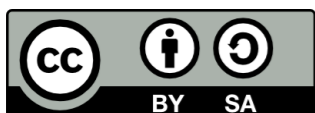
The emergence of the web has fundamentally affected most aspects of information communication, including scholarly communication. The immediacy that characterizes publishing information to the web, as well as accessing it, allows for a dramatic increase in the speed of dissemination of scholarly knowledge. But, the transition from a paper-based to a web-based scholarly communication system also poses challenges. In this paper, we focus on reference rot, the combination of link rot and content drift to which references to web resources included in Science, Technology, and Medicine (STM) articles are subject. We investigate the extent to which reference rot impacts the ability to revisit the web context that surrounds STM articles some time after their publication. We do so on the basis of a vast collection of articles from three corpora that span publication years 1997 to 2012. For over one million references to web resources extracted from over 3.5 million articles, we determine whether the HTTP URI is still responsive on the live web and whether web archives contain an archived snapshot representative of the state the referenced resource had at the time it was referenced. We observe that the fraction of articles containing references to web resources is growing steadily over time. We find one out of five STM articles suffering from reference rot, meaning it is

CrossMark

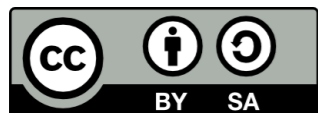
Subject Areas

- Archives
- Internet
- Extrapolation
- Communications
- Evolutionary immun...
- Computer and Infor...
- Ontologies
- Publication ethics

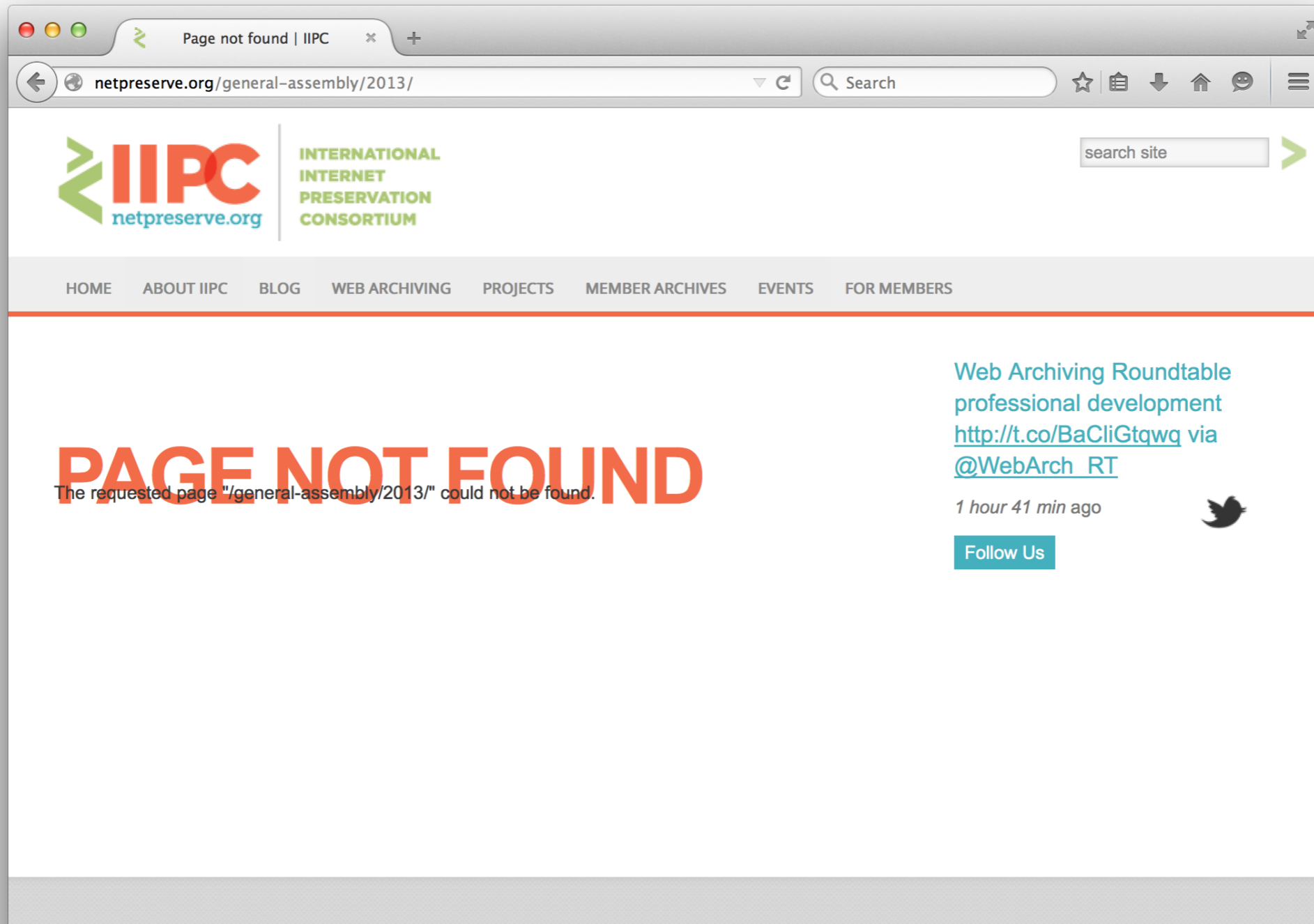
<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253>



# Reference Rot



# Link Rot



# “Entertaining” Link Rot



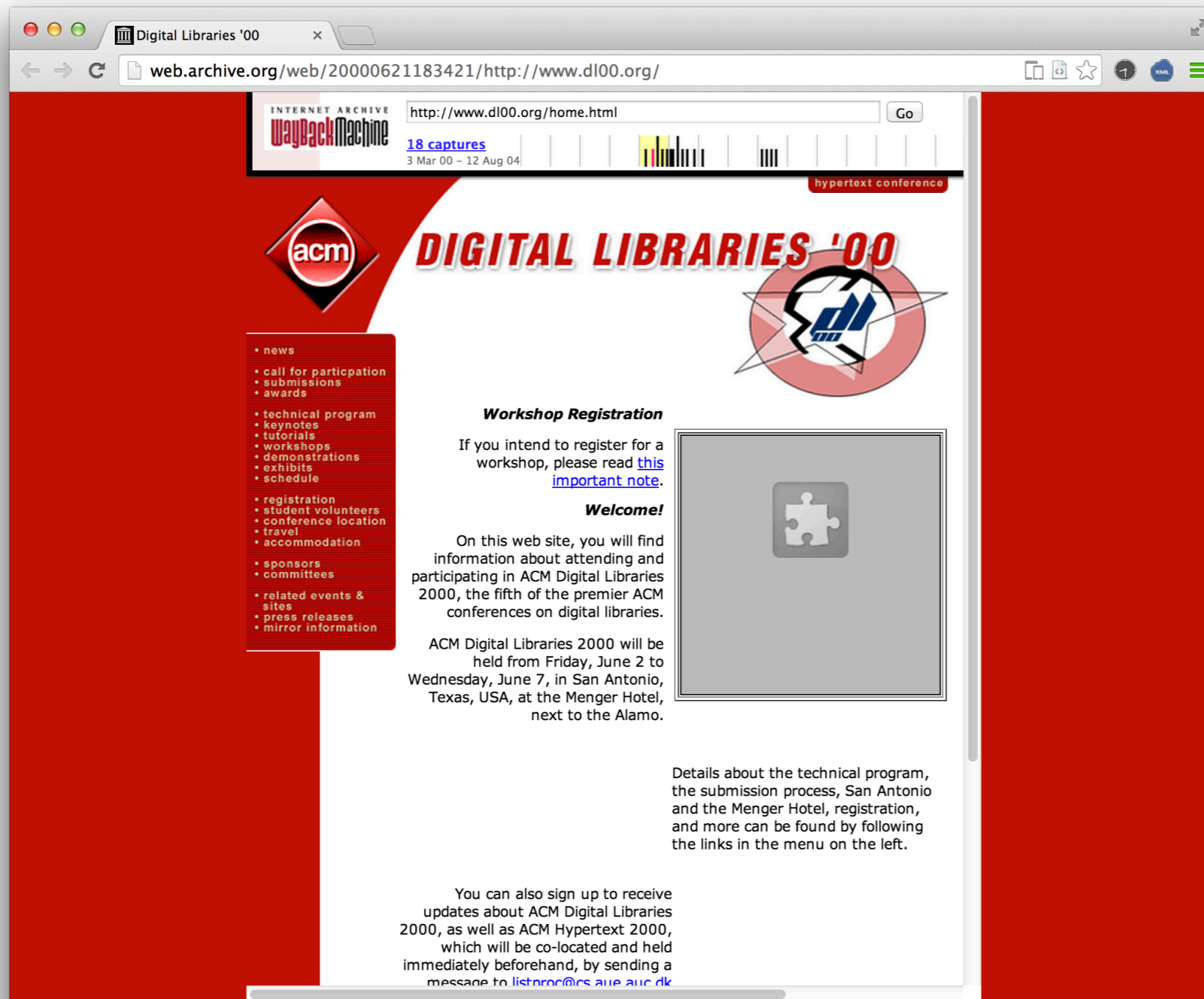
# Ubiquitous Link Rot



# Content Drift

<http://dl00.org>

2000





# Content Drift

<http://dl00.org>

**2004**

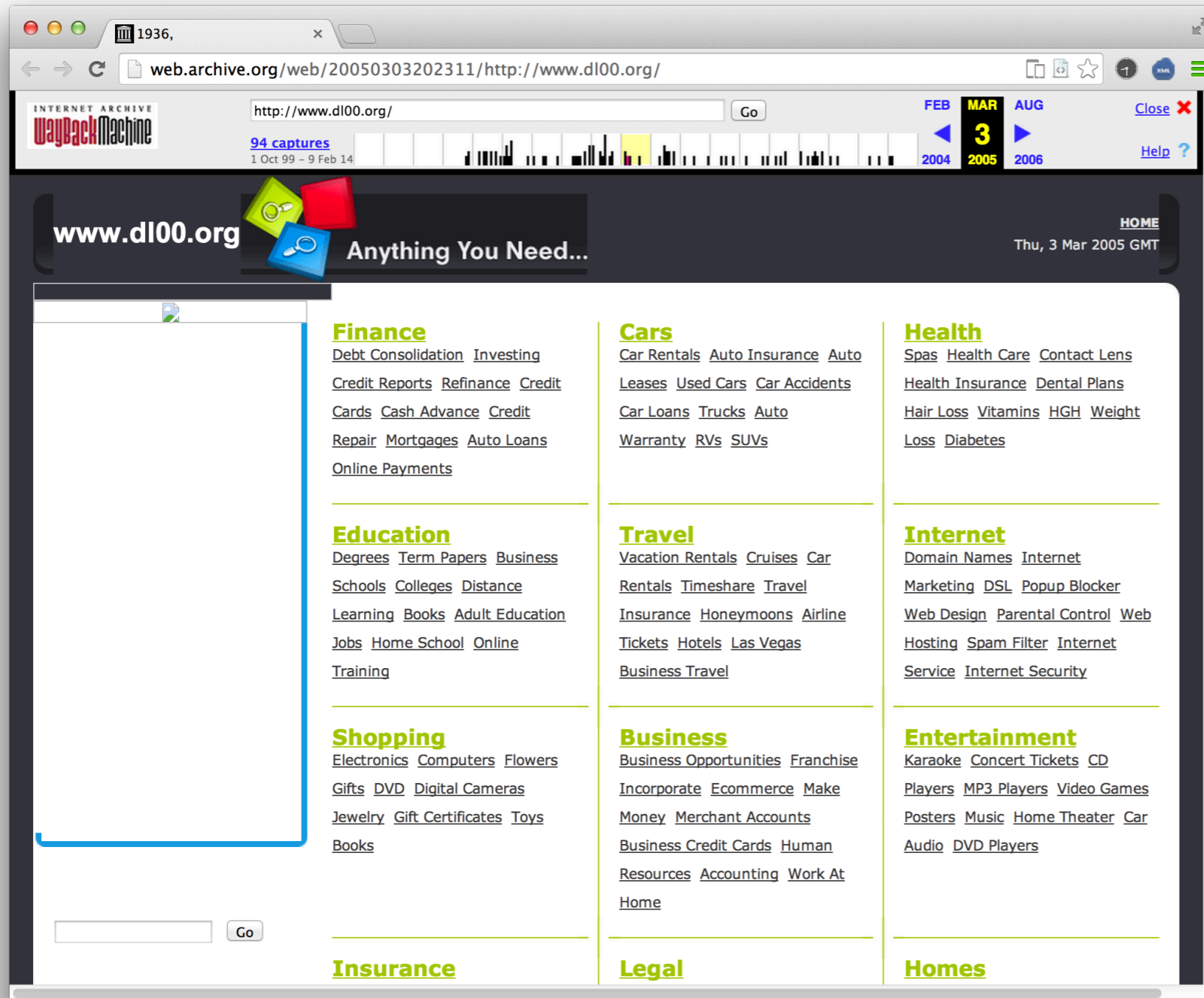
The screenshot shows a web browser window with the address bar displaying `web.archive.org/web/20040824033120/http://dl00.org/`. The page content includes a 'Wayback Machine' header with a calendar for August 2004, highlighting the 24th. Below the header is a 'directNIC' logo and a search bar. The main content area is organized into three columns of links:

- Left Column:**
  - Gibson
    - Gibson Appliance Part
    - Deborah Gibson
    - Passion
  - Los Angeles
    - Movie
    - Hard Drive
    - Musical Instrument
  - Photo
    - Acoustic Guitar
    - Science Fiction
    - Fan
- Middle Column:**
  - Gibson Guitar
    - Mel
    - Real Estate
    - William
  - St Louis
    - Stock Photography
    - Web Site
    - Auto Insurance
  - Bob Gibson
    - String
    - Civil War
    - Music
- Right Column:**
  - Gibson Les Paul
    - Guitar
    - Les Paul
    - Poster
  - Electric Guitar
    - Data Recovery
    - Bass Guitar
    - Accessory
  - Celebrity
    - Icon
    - Appliance Part
    - Electric

# Content Drift

http://dl00.org

2005



# Content Drift

<http://dl00.org>

2008

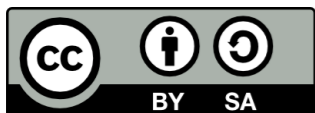
DL 2000 provides **project management, research and writing services** to produce detailed technical documents such as engineering and environmental reports, industry standards, government submissions, legislative reviews and technical papers.

We've worked extensively with industry to produce major reports and standards. Some of the highest profile documents we've produced include:

- The Basic Environmental Program and the detailed Environmental Operating Practices for the Upstream Petroleum Industry - Alberta Operations for the Canadian Association of Petroleum Producers
- Emergency Response Planning Guidelines for the Upstream Petroleum Industry for the Canadian Association of Petroleum Producers
- The Guide to Effective Public Involvement for the Canadian Association of Petroleum Producers
- Environmental impact statements
- Report of the Upstream Petroleum Industry Task Force on Safety and its Basic Safety Program for four major industry associations
- Guidelines for Designing Integrated Vegetation Management Plans for the Industrial Vegetation Management Association of Alberta
- Environmental annual reports
- Guidelines for the Handling of Naturally Occurring Radioactive Materials in Western Canada for the Canadian Association of Petroleum Producers
- Industry Standards and Good Practices for Vegetation Management for the Industrial Vegetation Management Association of Alberta
- Responses to Environment Canada's Green Plan and the Alberta Environmental Protection and Enhancement Act
- Policy Guides to assist the process of regulatory reform within the Ministry of Mines and the Ministry of Environment in Colombia

Member of [clean air trust](#)

[Advertising Enquiries](#)



# NYT Coverage

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.

[Enlarge This Image](#)



Stephan Savoia/Associated Press

Justice Samuel A. Alito Jr.

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.”

 FACEBOOK

 TWITTER

 GOOGLE+

 SAVE

 E-MAIL

 SHARE

 PRINT

 REPRINTS



Links in  
Supreme Court decisions:

- Link rot: **29%**
- Reference rot: **49%**

# Scholarly Communication

**D-Lib Magazine**  
**September 2004**

Volume 10 Number 9

ISSN 1082-9873

## **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@cscornell.edu>

[Simeon Warner](#)

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

## References

**!Exist**

Atkins, D. et al.. 2003. National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, *Revolutionizing Science and Engineering through Cyber-infrastructure*, <[http://www.communitytechnology.org/nsf\\_ci\\_report/](http://www.communitytechnology.org/nsf_ci_report/)>.

**Archived**

**Exist**

Brody, T., Kampa, S., Harnad, S., Carr, L. and Hitchcock, S. 2003. Digitometric Services for Open Archives Environments. In *Proceedings of European Conference on Digital Libraries 2003*, pages pp. 207-220, Trondheim, Norway. <<http://eprints.ecs.soton.ac.uk/archive/00007503/>>.

**Archived**

Frey, J., De Roure, D. and Carr, L. 2002. *Publication at Source: Scientific Communication from a Publication Web to a Data Grid*. <<http://eprints.ecs.soton.ac.uk/archive/00007852/>>.

Henry, G. 2003. On-line publishing in the 21-st Century: Challenges and Opportunities. *D-Lib Magazine*, Volume 9, Issue 10. <[doi:10.1045/october2003-henry](https://doi.org/10.1045/october2003-henry)>.

**!Exist**

Lynch, C. 2003. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL Bimonthly Report* 226. February 2003, <<http://www.arl.org/newsltr/226/ir.html>>.

**Archived**

**!Exist**

Payette, S., and Staples, T. 2002. The Mellon Fedora Project: Digital Library Architecture Meets XML and Web Services. *European Conference on Research and Advanced Technology for Digital Libraries*, Rome, Italy, September 2002. <<http://www.fedora.info/documents/ecdl2002final.pdf>>.

**!Archived**

Pöschl, U. 2004. Interactive Journal Concept for Improved Scientific Publishing and Quality Assurance. *Learned Information*, Volume 17, Number 2, pp 105-113. <[doi:10.1087/095315104322958481](https://doi.org/10.1087/095315104322958481)>.

Reich, V. and Rosenthal, D. 2001. LOCKSS: A Permanent Web Publishing and Access System. *D-Lib Magazine*, Volume 7, Issue 6. <[doi:10.1045/june2001-reich](https://doi.org/10.1045/june2001-reich)>.

**Exist**

Roosendaal, H., and Geurts, P. 1997. Forces and functions in scientific communication: an analysis of their interplay. *Cooperative Research Information Systems in Physics*, August 31 – September 4 1997, Oldenburg, Germany. <<http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html>>.

**Archived**

# Entrance Hiberlink

- These resources:
  - Are not necessarily under the custodianship of parties that care about long time integrity, access
  - Do not necessarily have the same sense of fixity like e.g., journal articles
- Links to these resources are subject to **Reference Rot**:
  - Link Rot: Link stops working e.g., HTTP 404
  - Content Drift: Linked content changes over time



# Quantifying Reference Rot



# Our Study

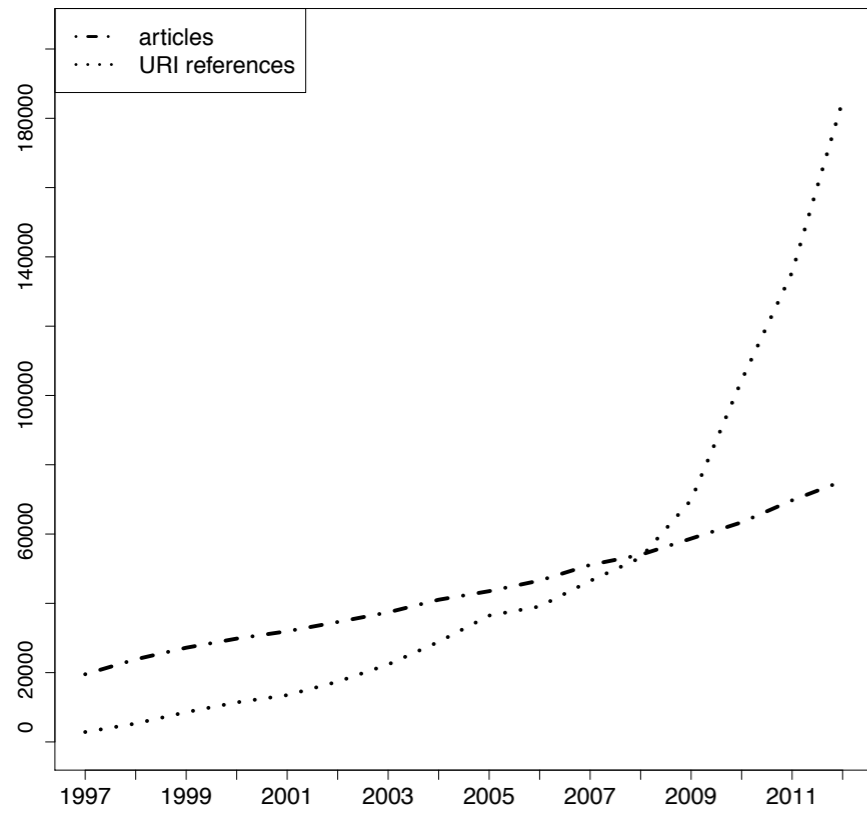
- Time frame of publications: Jan 1997 - Dec 2012
- Articles from arXiv, Elsevier, and PMC in XML and PDF format
  - Convert PDF to XML
  - Extract URIs to web at large resources
  - Store article's publication date
- URI live web test (trusted in 200 OK response)
- URI archive lookup via Memento infrastructure



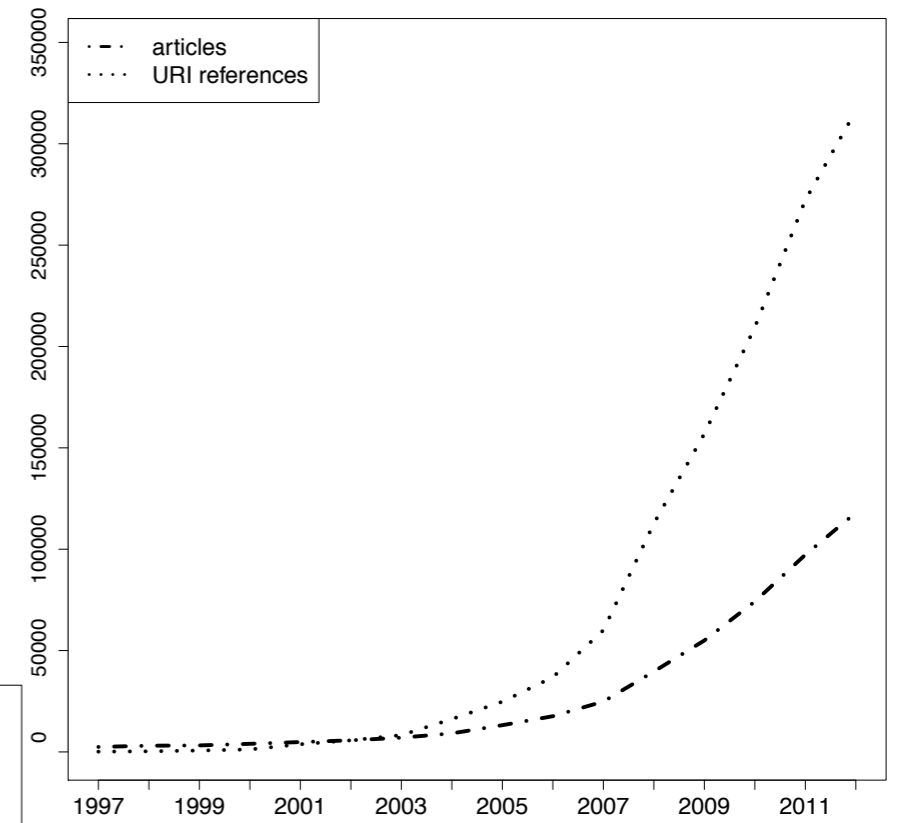
	<b>arXiv</b>	<b>Elsevier</b>	<b>PMC</b>
total articles	707,667	2,285,000	595,889
articles with HTTP references	142,134	94,645	156,160
amount of HTTP references	346,177	232,712	480,853

# Our Corpora

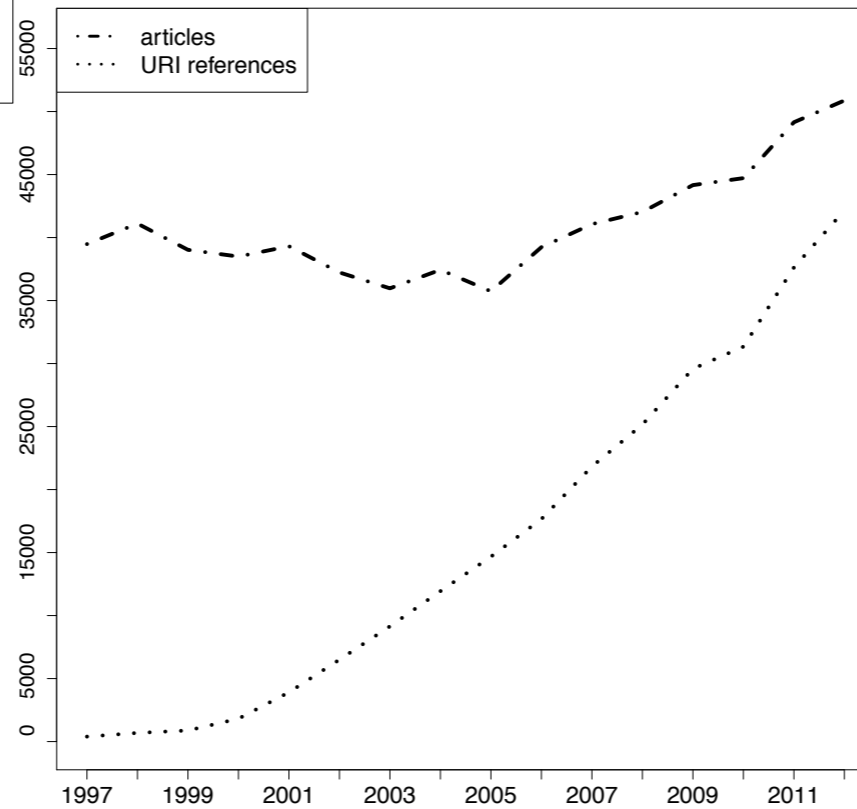
arXiv



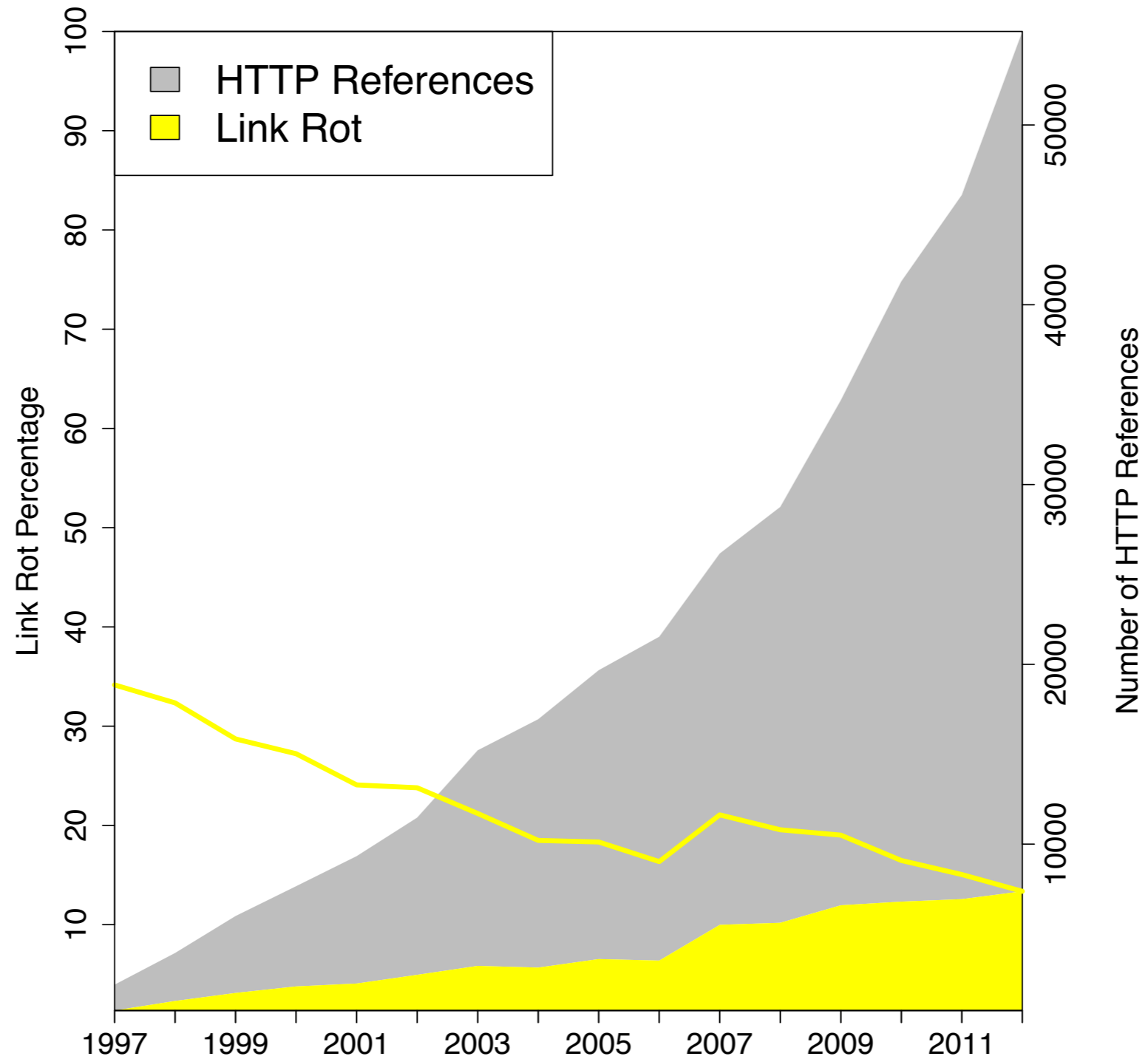
PMC



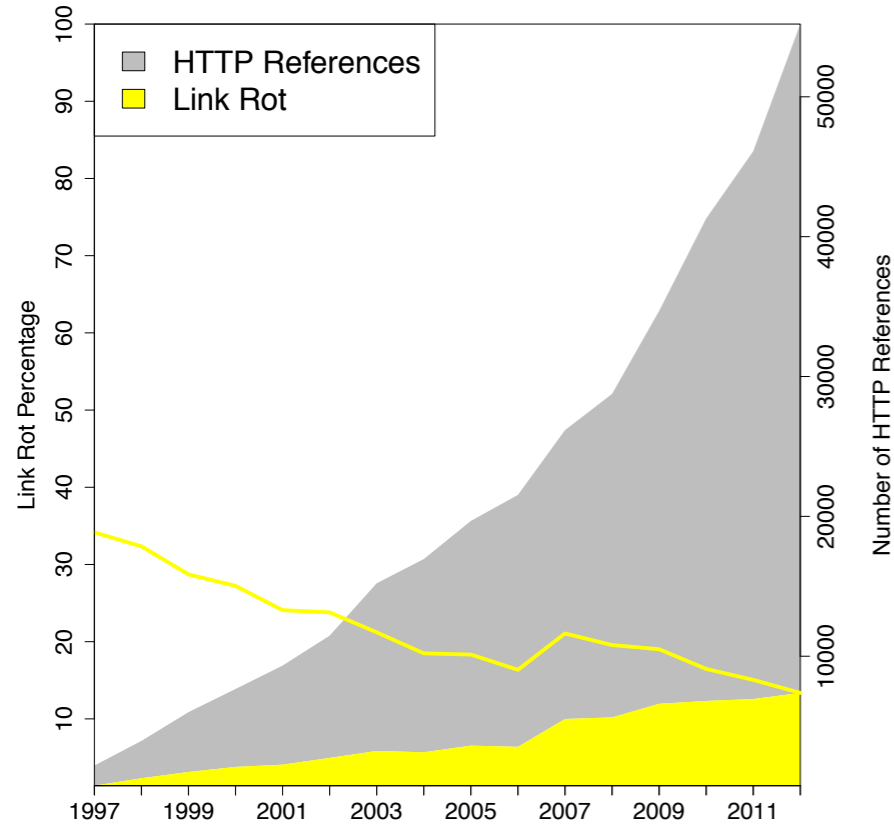
Elsevier



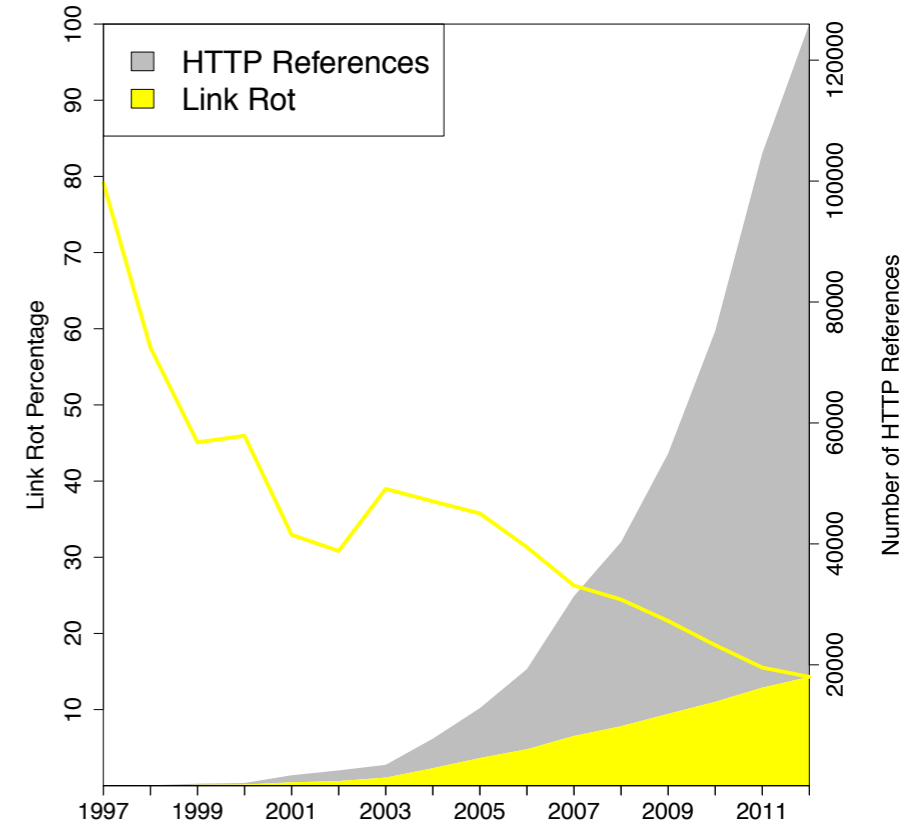
# Link Rot in arXiv



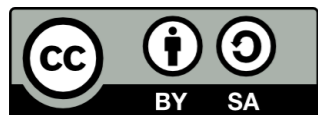
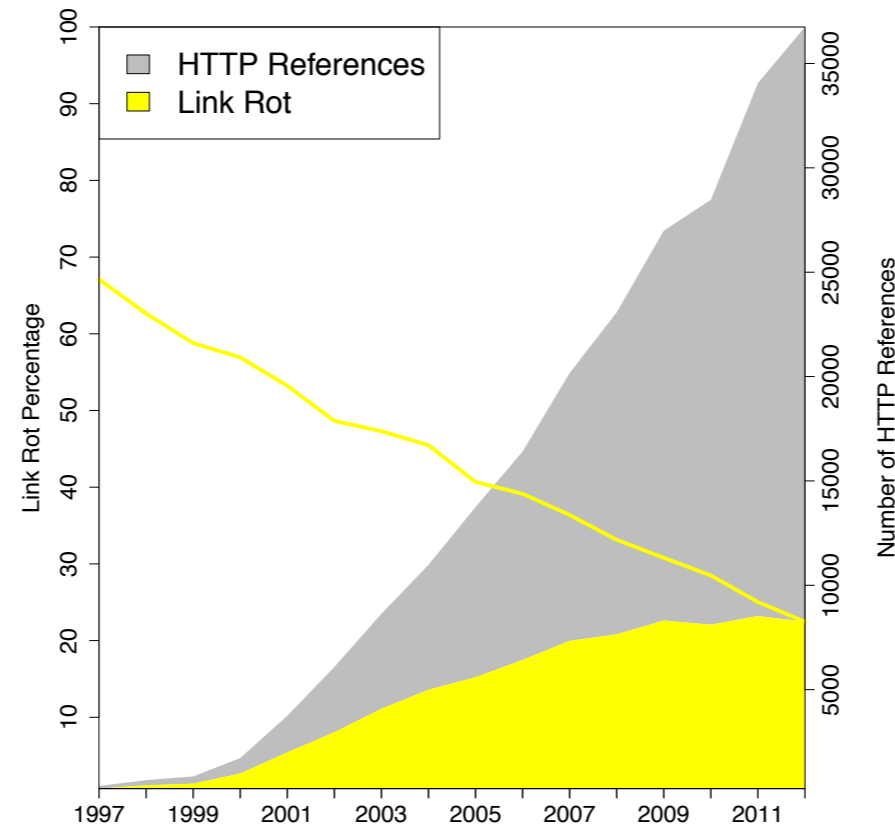
### arXiv



### PMC

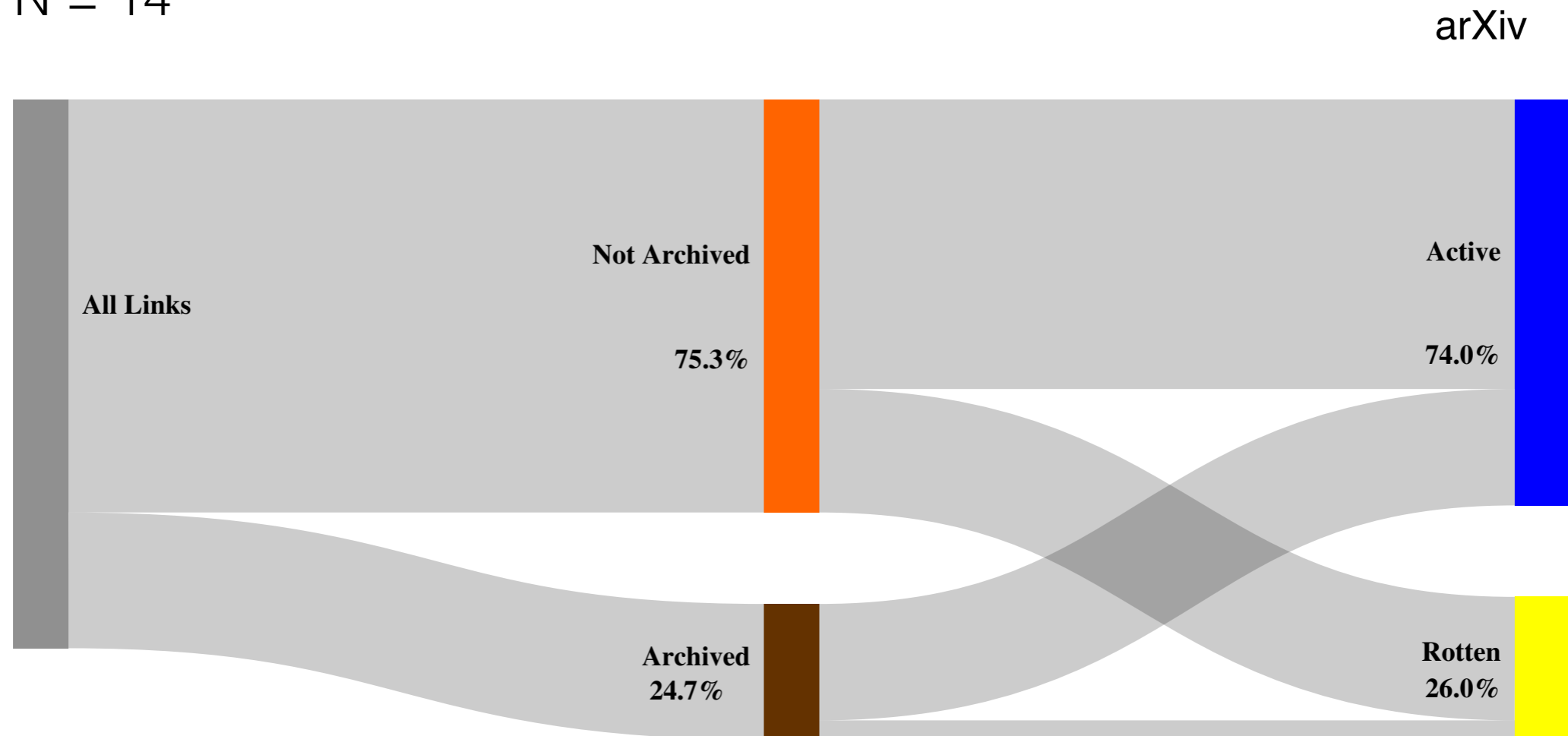


### Elsevier

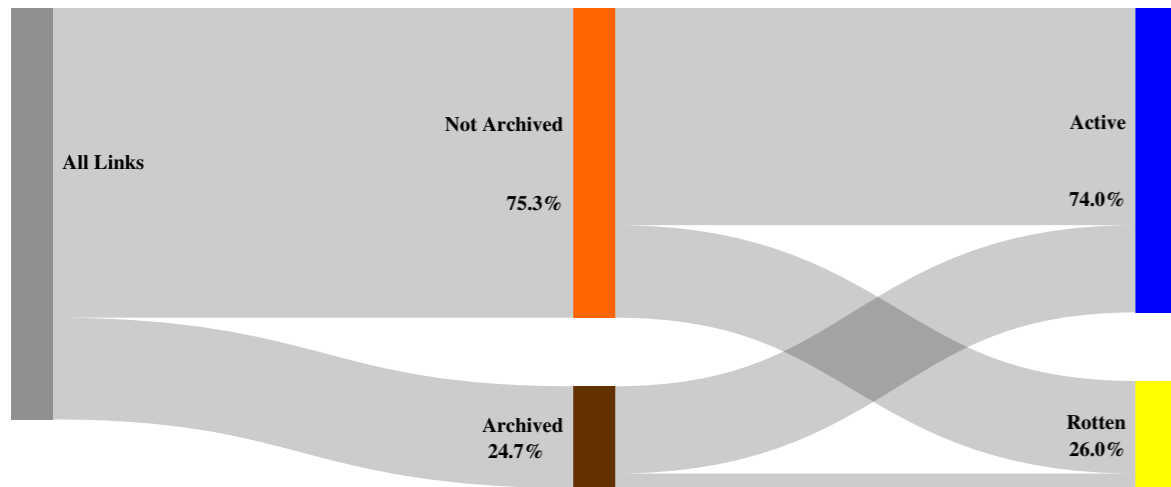


# Content Drift / Archival Status

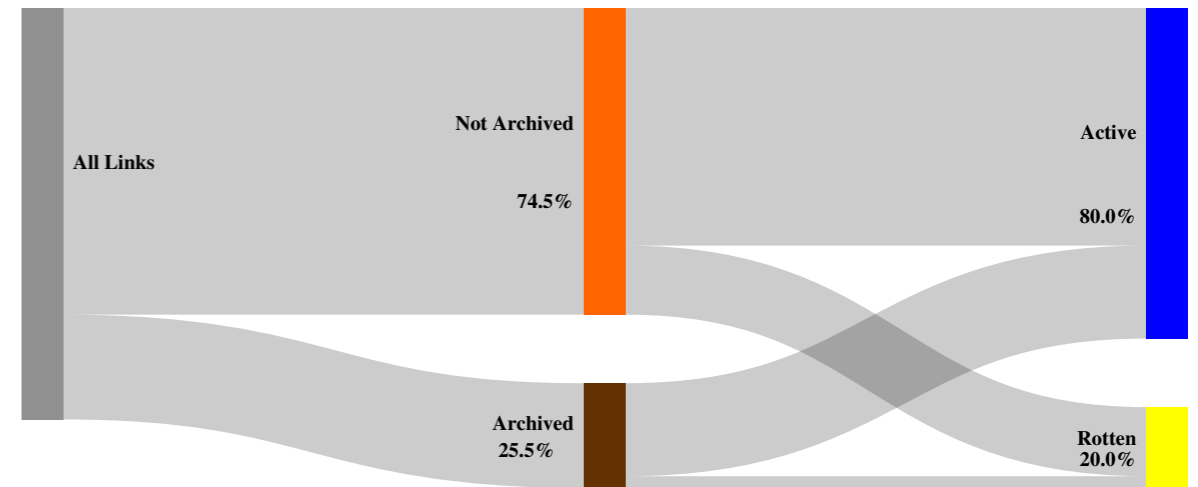
- Archival status used as proxy
- Availability of archived copy created within  $N$  days of article's publication
  - $N = 14$



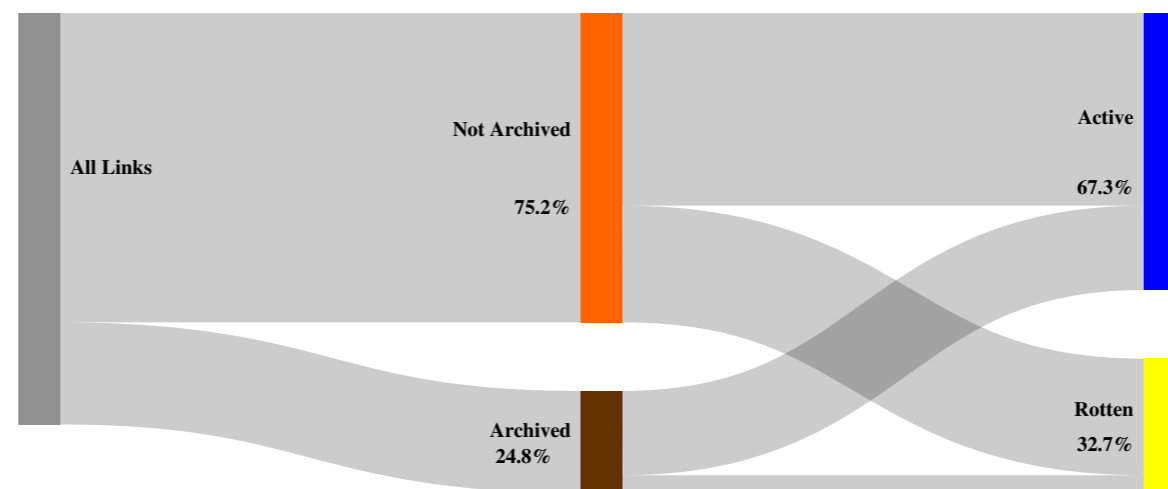
## arXiv



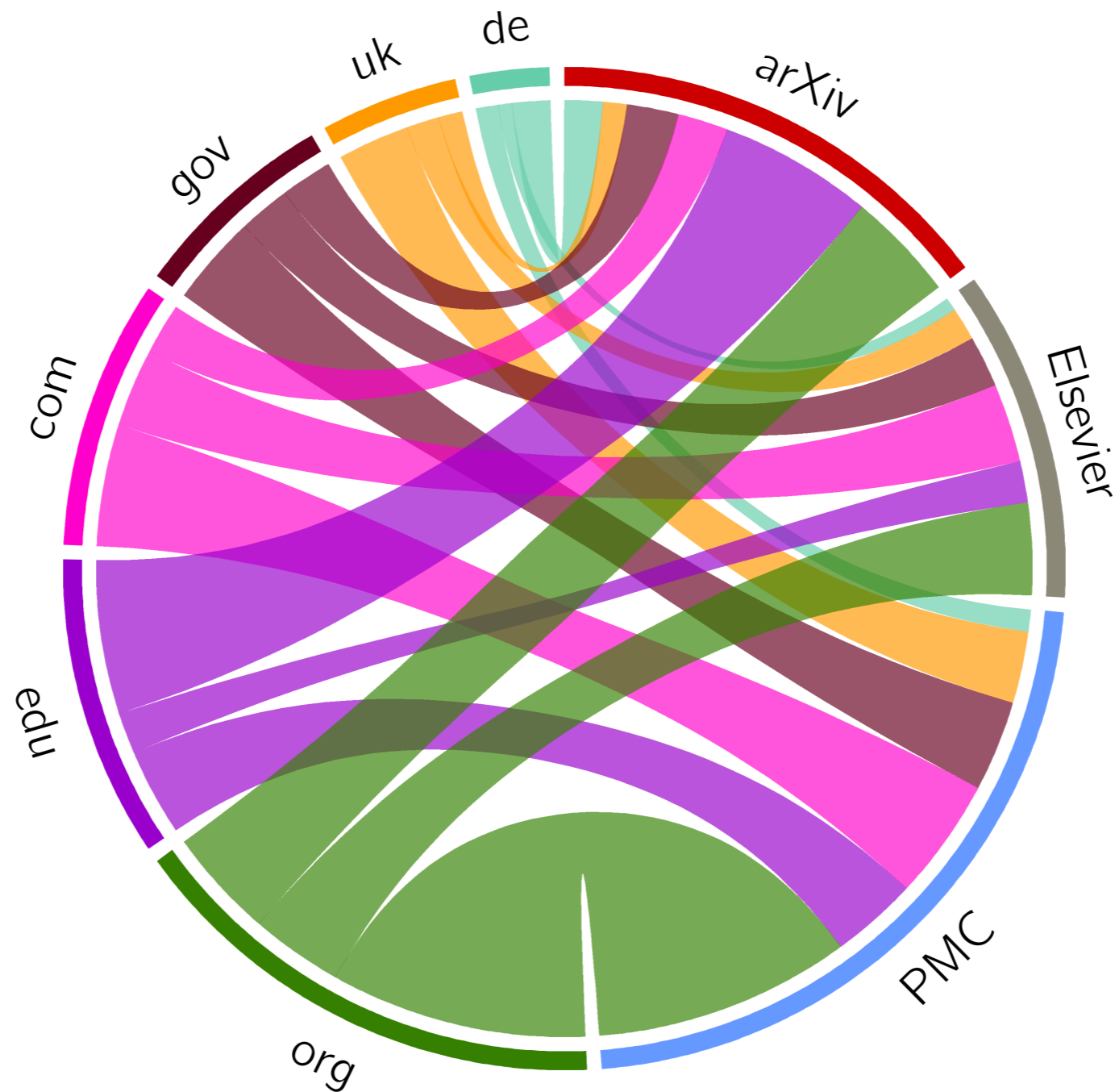
## PMC



## Elsevier

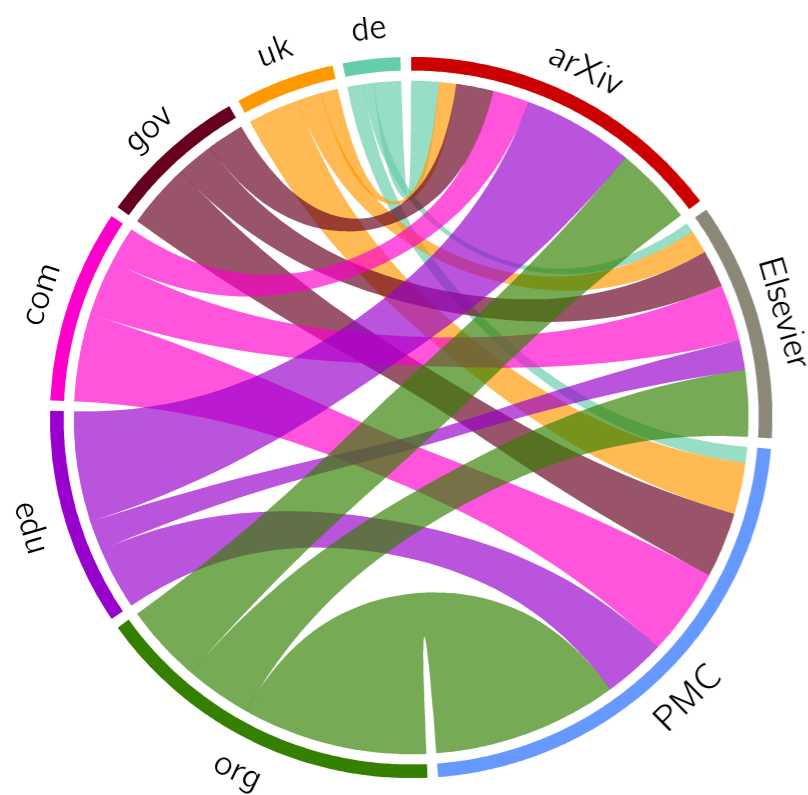


# Loss of Context

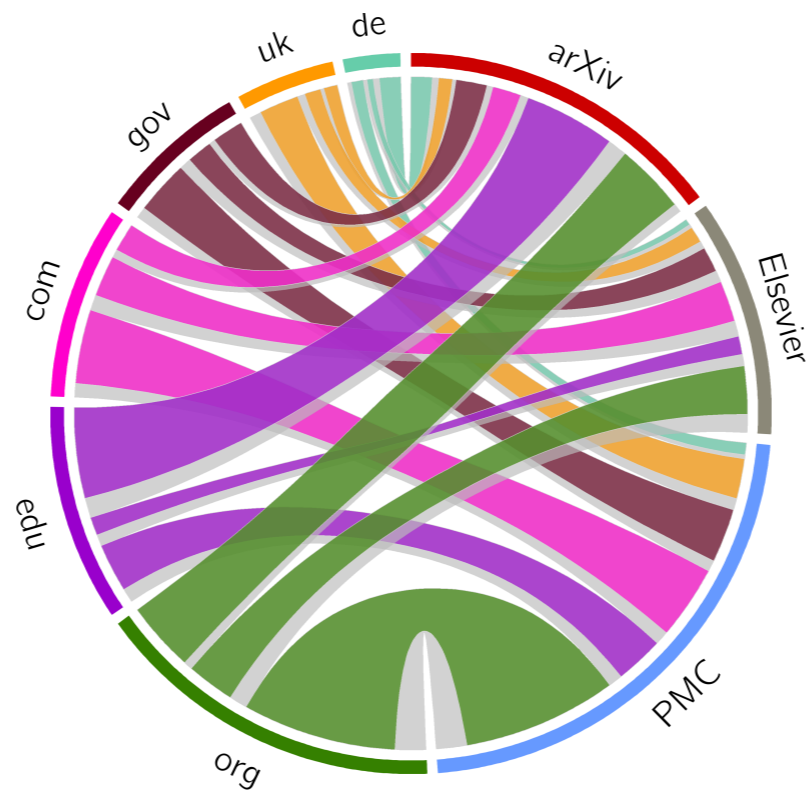


# Loss of Context

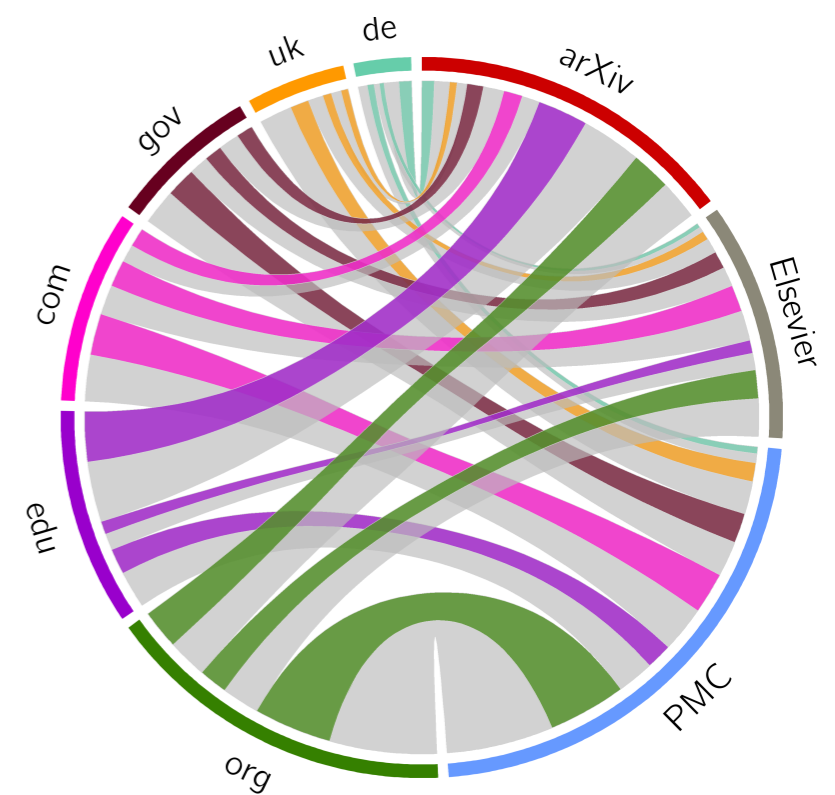
all links



active links

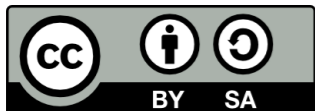


links archived  
(14 days)





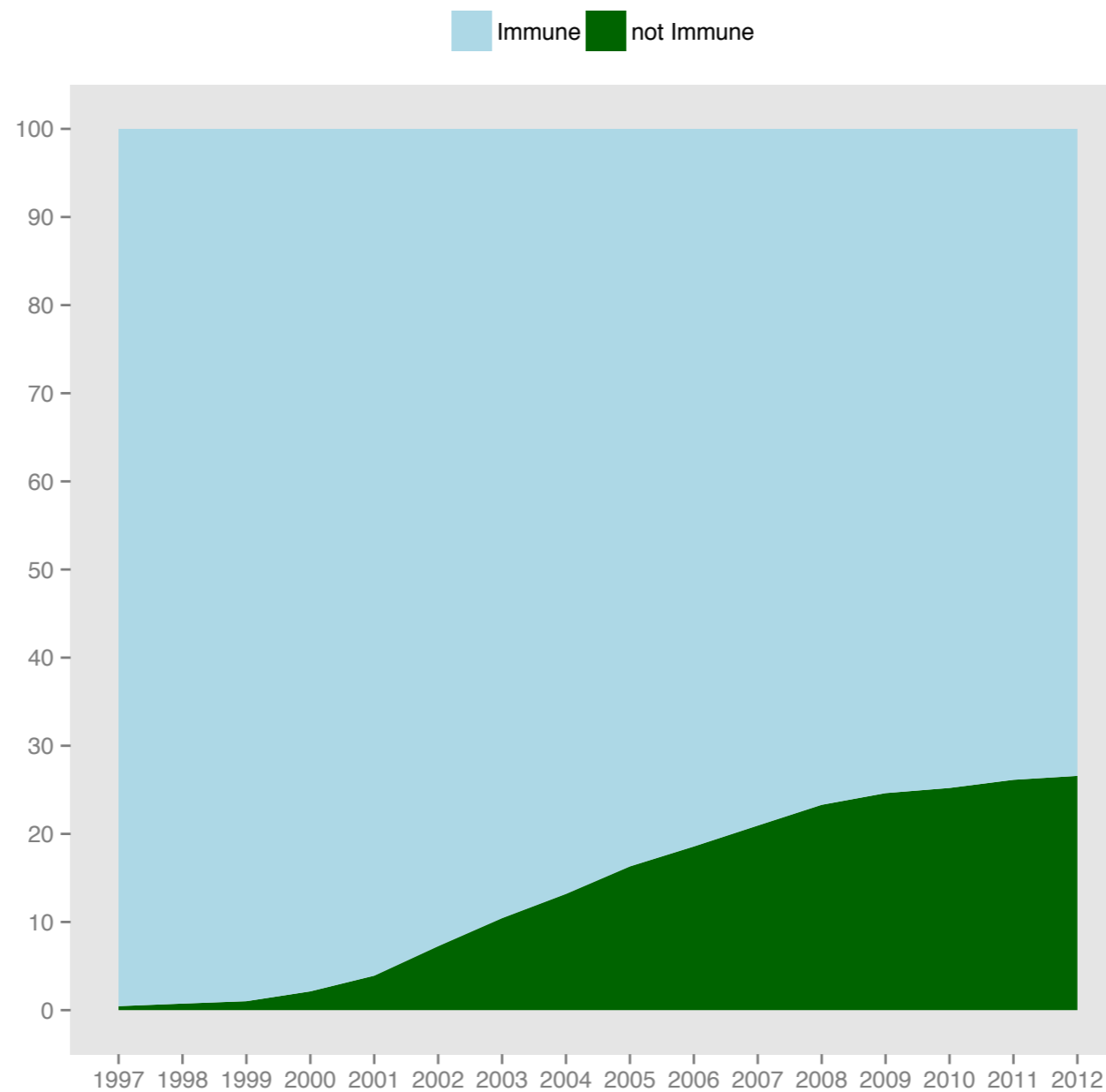
# STM Article Extrapolation



# STM Article Extrapolation

- Immune: article contains no URIs to web at large resources

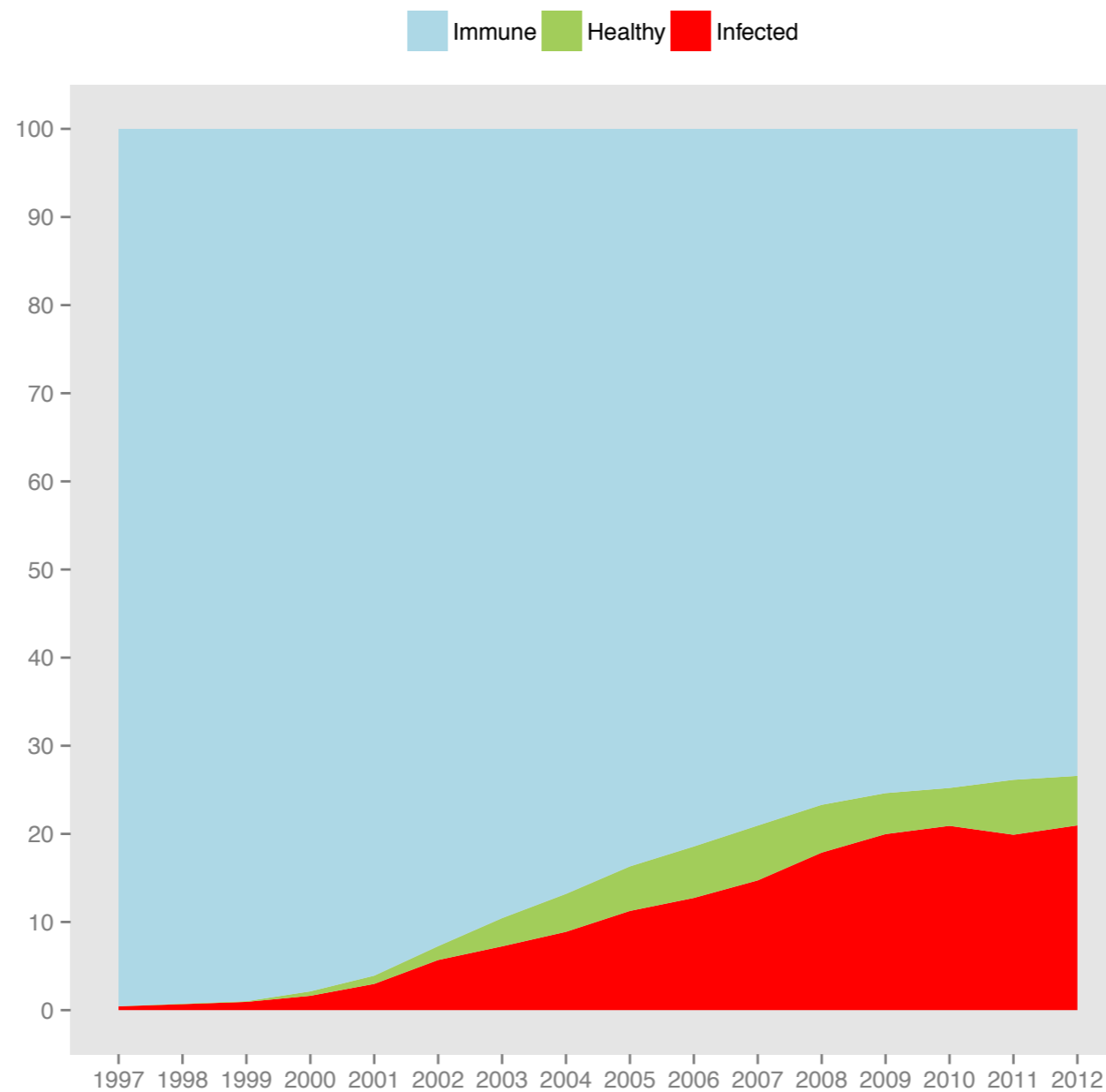
# Immune vs not Immune STM Articles



# STM Article Extrapolation

- Immune: article contains no URIs to web at large resources
- Healthy: **none** of the URIs to web at large resources suffer from reference rot
- Infected: **at least one** URI to web at large resources suffers from reference rot

# Immune, Healthy, Infected STM Articles



1/5 articles suffers  
from  
Reference Rot!

# An approach to solve Reference Rot

# Robust Links

1. Create **snapshot** of linked resources in a web archive when:

- drafting work
- submitting article
- publishing article
- aggregating article

**archive.is**  
webpage capture



**WebCite**

**perma.cc** ∞

# Robust Links

1. Create snapshot of linked resources in a web archive
2. Convey **creation date** of your web page in machine-actionable manner



# Page Creation Date

```
<!DOCTYPE html>  
<html>  
<head>  
<title> ... </title>  
<meta itemprop="datePublished" content="2015-02-18" />  
...  
</head>  
...  
</html>
```

POLITICS

# In Supreme Court Opinions, Web Links to Nowhere

SEPT. 23, 2013

Sidebar

By ADAM LIPTAK

Email

Share

Tweet

Save

More



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](#).

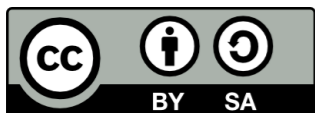
“Aren’t you glad you didn’t cite to this Web page?” it asks. “If you had, like Justice Alito did, the original content would have long since disappeared and someone else might have come along and purchased the domain in order to make a comment about the transience of linked information in the Internet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



```
<meta property="og:description" content="According to a new study, 49 percent of the hyperlinks in Supreme Court decisions no longer work." />
<meta property="article:published" itemprop="datePublished" content="2013-09-23" />
<meta property="article:section" itemprop="articleSection" content="Politics" />
```



# Robust Links

1. Create snapshot of linked resources in a web archive
2. Convey creation date of your web page in machine-actionable manner
3. **Decorate links** with datetime of linking and URI of archived snapshot, in addition to resource's original URI

<http://robustlinks.mementoweb.org/spec/>

# Link Decoration

`<a href="http://hiberlink.org/">http://hiberlink.org/</a>`

# Link Decoration

```
<a href="http://hiberlink.org/"
```

```
data-versionurl="http://archive.is/Bvq2v"
```

```
data-versiondate="2014-11-01">
```

```
http://hiberlink.org/</a>
```

# Robust Links - Reference List Demo (JavaScript)

Last updated (Date Published): February 18, 2015

## References are from the paper:

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou K., and Tobin, R. (2014) Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. PLoS ONE, 9(12): e115253. doi:10.1371/journal.pone.0115253 ; <http://dx.doi.org/10.1371/journal.pone.0115253>

## To see decorated links at work:

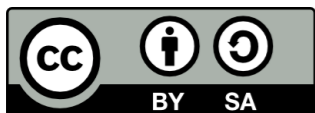
- Click on the ▼ arrow that is shown next to links in the Reference list to utilize link decorations
- [View the HTML source](#) to see the decorated links and the JavaScript that makes them operational

Reference List	
1.	Hiberlink <a href="http://hiberlink.org/">http://hiberlink.org/</a> ↗ Accessed: 1 Nov 2014
2.	Resolve a DOI Name <a href="http://dx.doi.org">http://dx.doi.org</a> Accessed: 1 Nov 2014
3.	LOCKSS <a href="http://lockss.org/">http://lockss.org/</a> ↗ Accessed: 1 Nov 2014
4.	CLOCKSS <a href="http://www.clockss.org/">http://www.clockss.org/</a> ↗ Accessed: 1 Nov 2014
5.	Portico - A Digital Preservation and Electronic Archiving Service <a href="http://www.portico.org/">http://www.portico.org/</a> ↗ Accessed: 1 Nov 2014
6.	The Keepers Registry <a href="http://thekeepers.org/">http://thekeepers.org/</a> ↗ Accessed: 1 Nov 2014
7.	Buckheit JB, Donoho DL (1995) Wavelab and reproducible research. In: Wavelets and Statistics, volume 103. pp. 55-81.
8.	Berners-Lee T (1998). Cool URIs don't change <a href="http://www.w3.org/Provider/Style/URI.html">http://www.w3.org/Provider/Style/URI.html</a> ↗ Accessed: 26 Nov 2014
9.	Koehler WC (2002) Web Page Change and Persistence - A Four-Year Longitudinal Study. Journal of the American Society for Information Science and Technology 53: 162-171.

[http://robustlinks.mementoweb.org/demo/uri\\_references\\_js.html](http://robustlinks.mementoweb.org/demo/uri_references_js.html)

Reference List		
1.	Hiberlink <a href="http://hiberlink.org/">http://hiberlink.org/</a> ↗	Accessed: 1 Nov 2014
2.	Resolve a DOI Name <a href="http://dx.doi.org">http://dx.doi.org</a>	Accessed: 1 Nov 2014
3.	LOCKSS <a href="http://lockss.org/">http://lockss.org/</a> ↗	Accessed: 1 Nov 2014
4.	CLOCKSS <a href="http://www.clockss.org/">http://www.clockss.org/</a> ↗	Accessed: 1 Nov 2014
5.	Portico - A Digital Preservation and Electronic Archiving Service <a href="http://www.portico.org/">http://www.portico.org/</a> ↗	Accessed: 1 Nov 2014
6.	The Keepers Registry <a href="http://thekeepers.org/">http://thekeepers.org/</a> ↗	Accessed: 1 Nov 2014
7.	Buckheit JB, Donoho DL	Robust Links
8.	Berners-Lee T (1998). C	Get near page date 2015-02-18
9.	Koehler WC (2002) Web Information Science and	Get near link date 2014-11-01
10.	The Chesapeake Digital Chesapeake Digital Pres	Get from archive.today
11.	Zittrain J, Albert K, Lessig L (2014) Perma: Scoping and addressing the problem of link and reference rot in legal Law Review Forum 127.	

[http://robustlinks.mementoweb.org/demo/uri\\_references\\_js.html](http://robustlinks.mementoweb.org/demo/uri_references_js.html)



# Reference Rot and Link Decoration

**Martin Klein**

UCLA

[martinklein0815@gmail.com](mailto:martinklein0815@gmail.com)

[@mart1nkle1n](#)

